Journal of Computer-Aided Molecular Design, 6 (1992) 487–504
ESCOM

J-CAMD 176

QSAR of conformationally flexible molecules: Comparative molecular field analysis of protein-tyrosine kinase inhibitors

Marc C. Nicklaus*, George W.A. Milne and Terrence R. Burke, Jr.

Laboratory of Medicinal Chemistry, Developmental Therapeutics Program, Division of Cancer Treatment, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, U.S.A.

Received 3 March 1992
Accepted 16 July 1992

Key words: CoMFA; Styrene derivatives; Prediction of bioactive conformation

SUMMARY

Comparative Molecular Field Analysis (CoMFA) has been applied to a study of quantitative structure–activity relationships (QSAR) of conformationally flexible molecules. The relationship between three-dimensional structure and activity of 20 styrene derivatives which inhibit protein-tyrosine kinase was determined. A technique was developed that allows accurate prediction of the inhibitory activity of these molecules and identification in each case of the active conformation. The problem of multiple energetically acceptable conformations was approached in an iterative procedure. Use was made of the varying degrees of symmetry among the molecules. First, CoMFA QSAR models were developed using only those compounds that possess a symmetrical substituent pattern on the phenyl ring. These CoMFA models were then used to select the active conformers of the less symmetrical compounds in the set. Allowing multiple conformers for each compound in the dataset yielded higher crossvalidated r^2 values and better predictivity of the QSAR models. Different probe atoms (C^+, O^-, neutral C) were explored, the O^- probe atom exhibiting the highest selectivity in the conformer selection process.

INTRODUCTION

When designing compounds that will bind to specific receptor sites on an enzyme whose detailed structure is not known, it is usualy necessary to carry out an analysis of the relationships between structure and activity for those 'standard' compounds for which biological data are already available. When the standard compounds are capable of assuming different conformations, this analysis becomes difficult.

Protein-tyrosine specific kinases (PTK) are an important class of phosphoryl transfer enzymes which mediate signal transduction processes associated with normal cellular growth and differentiation [1, 2]. These enzymes also play key roles in certain neoplastic processes [3–5] and have been

* To whom correspondence should be addressed.

0920-654X/$ 5.00

associated with the etiology of a number of human cancers [6,7]. Agents which specifically inhibit the function of PTKs are logical targets in the development of new anticancer chemotherapeutics [8,9], and a number of such compounds have been prepared [10–13]. The lack of 3D structural information for PTKs has limited the design of new inhibitors, and frequently new analogs are either synthetic extensions of structural motifs derived from the screening of natural products [14, 15] or extrapolations by 'well educated guesses' from existing molecules with known biological activities [16]. The refinement of quantitative structure–activity relationship (QSAR) methodologies to enhance this latter process is therefore of particular value in the development of inhibitors for these enzymes.

Comparative Molecular Field Analysis (CoMFA) [17,18] is a method which uses multivariate regression techniques to seek relationships between properties such as biological activity and various 3D structural features such as molecular shape and charge distribution. A number of applications of CoMFA in medicinal chemistry have been reported [18–30]. However, the methodology in its present state offers no explicit provision for management of different conformations of molecules with rotatable bonds [28]. No attempt at an explicit solution of this conformational problem has to our knowledge been published so far. Studies that involved the application of CoMFA have either been conducted on essentially rigid [17,19] or conformationally unambiguous [28] molecules, or the lowest energy conformation of each compound was selected in the case of flexible molecules [25–27], or fixed conformations as determined by X-ray crystallography were used [23,24]. A recent paper by Greco et al. [29] illuminates the difficulties encountered with external criteria for the selection of conformers in a CoMFA of flexible molecules. Recently, Kim and Martin [30] have used averaging to derive predicted values for compounds with multiple conformers in a CoMFA study on dissociation constants. This approach, however, does not seem to be generally applicable to ligand–receptor interactions where the assumption is made in the standard case that just one conformation of a flexible molecule will fit optimally in the receptor site.

This paper reports the application of CoMFA in a study of the relationship between detailed structural changes in a series of conformationally flexible styrene derivatives and the PTK inhibitory activity exhibited by these compounds. Useful correlations have been derived from this study and a promising approach to the conformational problem is reported. A procedure is presented that permits the prediction of the biological activity and the identification of the active conformer of conformationally flexible molecules not previously contained in the series.

MATERIALS AND METHODS

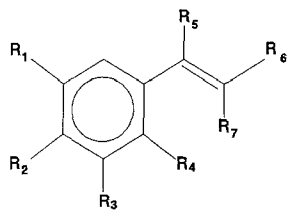
Twenty styrene derivatives with associated PTK inhibitory activity as reported by Yaish et al. [32] and identified in Table 1 were used in this study. The inhibition of PTK-mediated autophosphorylation is recorded in Table 1 in terms of K_i (μmol) and pK_i values, in epidermal growth factor receptor (EGFR) derived from A-431 cells.

3D models of all the molecules shown in Table 1 were generated with the Molecular Mechanics program SYBYL [33], run on a Silicon Graphics 4D/70G workstation. A benzene fragment was used as the nucleus for all molecular models. Conformational energy calculations and minimizations were carried out using the following modeling parameters and program options: the Tripos 5.2 force field [34] was used with a distance-dependent dielectric function and the standard Tripos parameter set. Non-bonded interactions were cut off at 12 Å. Electrostatic interactions were in-

cluded, and the Gasteiger method [35,36] was used for calculation of the partial atomic charges. The non-bonded list was updated at every tenth step. Minimization was carried out using the BFGS method [37], which was allowed to iterate through 100 steps or until convergence was achieved (defined as an energy change $\leq$ 0.05 kcal/mol between subsequent minimization steps).

All these molecules possess at least one easily rotatable bond, between the phenyl ring and the vinylic double bond. In general, there are two orientations of this bond which correspond to local energy minima. These are the two possible orientations of the aromatic ring coplanar with the vinylic double bond. Steric interactions lead to an angular deviation from coplanarity in these structures typically between 20° and 40°. Structures such as **1**, **4** and **9** (Table 1), in which R_6 or R_7 are substituents other than the rotationally symmetrical groups H or CN, possess an additional conformational degree of freedom given by the rotation of those substituents. For every molecule, rotatable bonds were examined by systematic conformational searching techniques, employing rigid rota-

TABLE 1
SET OF STYRENE DERIVATIVES USED IN THIS STUDY



Compound	R_1	R_2	R_3	R_4	R_5	R_6	R_7	K_i[a] (μmol)	pK_i
1	OH	OH	H	H	H	$CSNH_2$	CN	0.85	6.071
2	OH	OH	OH	H	H	CN	CN	1	6.000
3	OH	OH	OCH_3	H	H	CN	CN	2	5.699
4	OH	OH	H	H	H	$CONH_2$	CN	2.3	5.638
5	OH	OH	OH	H	OH	CN	CN	3.3	5.481
6	OH	H	OH	H	H	CN	CN	10	5.000
7	OH	OH	H	H	H	CN	CN	11	4.959
8	OH	OH	H	H	H	COOH	CN	18	4.745
9	OH	H	H	OH	H	CN	COOH	24	4.620
10	H	CHO	H	H	H	COOH	CN	47	4.328
11	OCH_3	OH	H	H	H	CN	CN	67	4.174
12	H	OH	H	H	OH	CN	CN	77	4.114
13	OH	NO_2	H	H	H	CN	CN	150	3.824
14	H	OH	H	H	H	COOH	CN	166	3.780
15	OCH_3	OH	OCH_3	H	H	COOH	CN	233	3.633
16	OCH_3	OH	H	H	H	COOH	CN	267	3.573
17	H	OCH_3	H	H	H	COOH	CN	833	3.079
18	H	F	H	H	H	COOH	CN	833	3.079
19	H	OH	H	H	H	CN	H	1333	3.125
20	OH	H	H	H	H,H[b]	CN	CN	2500	2.602

[a] Inhibition of PTK-mediated autophosphorylation in EGFR derived from A-431 cells (see Ref. 32).
[b] The vinylic double bond has been replaced by a single bond.

tion around those bonds with a step size of 10°–30°. Evaluation of the resulting energy surface yielded 2 or 4 local minimum energy conformations. Each local minimum structure obtained in this way was energy minimized before it was entered in the CoMFA database. The maximal energy difference between the individual conformers of a compound was typically $\leq$ 4 kcal/mol. The energy barriers of the torsions examined in the conformational searching were calculated to be typically between 2.5 and 5 kcal/mol.

A closer inspection of the molecules reveals that the two possible orientations of the phenyl ring with respect to the vinylic bond are not really different in all compounds. In 10 cases (compounds **2, 5, 6, 10, 12, 14, 15, 17, 18** and **19**), the phenyl ring is symmetrically substituted and the two conformations therefore are identical with respect to the substituent distribution. These molecules will therefore be termed '(more) symmetrical' (see Fig. 1). However, molecular modeling (since it simulates a structure at 0 K) assigns fixed positions to all atoms in a molecule. For some of the symmetrical compounds (e.g. **5** or **12**), this leads the conformational search process to deliver structures differing only in the ring hydroxyl H orientation as distinct conformations. The high ease of rotation of hydroxyl hydrogens at ambient temperature (unless intramolecularly hydrogen-bonded) [38] would not usually lead one to regard those structures as truly different conformers. However, none of those conformations were excluded from the computation at this point because they each possessed quite different fields and none of the conformations could, a priori, be eliminated on the ground of their relative conformational energies, which are nearly identical (energy differences < 0.4 kcal/mol). An attempt to deal with the problem of fixed hydrogen orientations in the context of CoMFA will be presented in the next section.

Once minimized conformations of all the structures are in hand, the next step in the CoMFA analysis is to establish a mutual spatial alignment of the molecules within the series. This can be done by computing the fields associated with each of the molecules and then minimizing the RMS differences between them by adjusting their positions (using the 'Field fit' technique in SYBYL). However, with the given structures, belonging all to the same structural family, this was felt to be

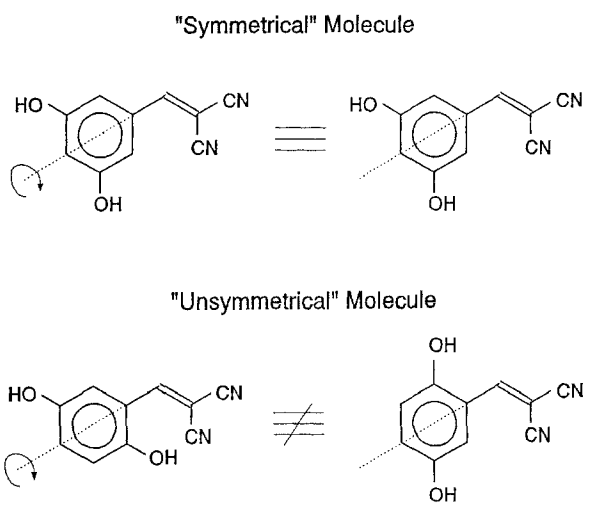


Fig. 1. Example for 'symmetrical'/'unsymmetrical' molecules.

unnecessarily complicated and a process which seemed more intuitive was adopted, consisting of simply superimposing the 8-carbon styrene residues.

The non-bonded potential energy fields associated with any structure are determined in CoMFA by first placing the structure within a 3D lattice and then computing the steric (Lennard–Jones) and electrostatic interactions with the molecule that are experienced by a probe atom (e.g. O^- or C^+) at all of the lattice points around the molecule. A lattice grid spacing of 2 Å in the x, y and z directions was used. These data are built into a CoMFA table which has 1 row per 'standard' compound, typically one column for the biological data (e.g. $-\log K_i$ values), and hundreds, or thousands, of columns to carry all the CoMFA energy field data. The statistical task at this stage is to establish by regression analysis over the whole set of compounds a relationship between these variables on the one hand, and the biological activity on the other. The biological data were entered in the CoMFA data tables as $-\log K_i$ values.

Because of the very large number of independent variables in this system, the standard techniques of multiple regression cannot be applied directly and Partial Least Squares (PLS) analysis [39–41] is employed to derive linear equations from the data. The quality of any model derived can be determined by a crossvalidation technique, sometimes referred to as the 'leave-one-out' method. It measures the true predictive capability of the model by removing a subset ('crossvalidation group') of one or more structures from the set of the standard compounds (the 'training set'), re-entering them as unknowns and predicting their activity. The sum of the squared differences between predicted and actual activities is compared to the standard deviation of the actual activities, to yield a 'crossvalidated r^2' as a measure of predictivity [18,42]. This value can range from negative numbers to a maximum of $+1.0$. A crossvalidated r^2 of $+1.0$ corresponds to a 'perfect model', i.e. all predictions were completely accurate. A value of 0 indicates that no statistical relationship has been detected between the CoMFA fields and the activities, and the predictions are no better than random numbers whose average is equal to the mean of the measured activities. Negative values result from an inverse relationship between actual and predicted activity, and can be regarded as 'worse than no model at all'. In all crossvalidation runs performed in the course of this study, the number of crossvalidation groups was chosen equal to the number of structures in the CoMFA table, i.e. the activity of each structure was predicted individually from all other structures. This guarantees reproducibility and comparability of subsequent crossvalidation runs*. The standard deviation threshold for exclusion of columns from the PLS analysis (MINIMUM_SIGMA) was set to 0.5 for crossvalidation runs, and to 0.0 for the final, non-crossvalidated PLS analyses.

CoMFA models can be built using a variable number of components in the linear equations derived by the PLS analysis. This number corresponds to the level of detail at which the PLS analysis has been conducted on the molecular fields, and is analogous to the number of coefficients in conventional multiple regression [42]. Up to $n-1$ components can be used, with n the number of structures in the CoMFA dataset. In the crossvalidation runs, a crossvalidated r^2 value is calculated for each value μ in the specified range of number of components ($\mu = 1...m$, $m \leq n-1$). CoMFA QSAR models that allow the prediction of the activity of novel compounds are derived from

*If the number of crossvalidation groups chosen is smaller than the number of structures in the CoMFA dataset, the distribution of the structures among those crossvalidation groups is done by CoMFA in a random fashion that cannot be controlled.

(final) non-crossvalidated PLS runs. For those models, one specific number of components had to be chosen. A number was usually chosen that was lower than the one yielding the highest crossvalidated r^2, since the number of components at which the usually steep initial increase of the crossvalidated r^2 starts to level off has proven to yield better predictive CoMFA models*. Higher numbers of components tend to 'over-fit' the CoMFA model to irrelevant minutiae of the molecules in the dataset and may lead to less accurate predictions [43].

The problem of multiple conformations was approached by first developing a CoMFA model with the symmetrical compounds that have the smaller degree of conformational ambiguity (such as **2, 6** or **19**). This model was then used to select the active conformers of the less symmetrical compounds in the set in a stepwise procedure. This procedure consisted of only including in each step those compounds whose biological activity had been predicted within a given limit of accuracy by the previous step CoMFA model. Different probe atoms were investigated as to their influence on the maximum attainable correlation coefficients as well as to their selectivity for conformers in the stepwise selection process.

RESULTS

1. C^+ probe atom

1.1. Single-conformer CoMFA models

Table 2 shows CoMFA models derived from various subsets of compounds **1–20**, using an sp^3-hybridized C^+ atom (charge +1.0) as the probe (the default probe atom offered by SYBYL). A selection was sought of exactly one conformer per compound with the objective of obtaining a CoMFA QSAR model with the highest possible crossvalidated r^2.

In a preliminary step, the two lowest energy conformers of each of the 10 symmetrical molecules were submitted to CoMFA. In each case, CoMFA predicted a K_i value for the structure. These predicted K_i values were compared to the measured values and if they differed by more than $\approx$30%, that structure was removed from the dataset. In this preliminary pass, only 3 structures (**2, 5** and **19**) survived and they comprised CoMFA dataset No. 1, shown in Table 2. This new set of 3 structures was re-analyzed and a maximum crossvalidated r^2 of −0.534 was obtained. This value, which would normally indicate a very bad CoMFA model, has no profound physical meaning at this stage, and the number of structures in the dataset is far too small for a generally useful QSAR model in any case. The exact procedure for derivation of this initial CoMFA model does not seem to be of paramount importance, since the model is used only as a baseline from which the subsequent steps can be evaluated. This model was next used to predict the K_i values of the remaining 2 most symmetrical compounds (**6** and **12**). For both compounds, the conformer whose predicted K_i was closest to the actual value was added to the dataset (yielding CoMFA dataset 2). This new dataset's maximum crossvalidated r^2 was −0.024. These 5 compounds were then recycled through the process iteratively to produce the CoMFA datasets, 3, 4, and 5. For CoMFA

*As an example, the crossvalidated r^2 values are listed for CoMFA dataset 4 (Table 2) as a function of the number of components ranging from 1 through 8: 0.399, 0.610, 0.664, 0.698, 0.703, 0.713, 0.713, 0.706. As can be seen, the maximum is reached for six components; clearly, however, the crossvalidated r^2 starts to level off at four components.

datasets 1–3, only symmetrical molecules were used, and only in CoMFA datasets 4 and 5 were the less symmetrical molecules added.

It can be seen that through datasets 1–3, the maximum crossvalidated r^2 remains quite low. However, it increases substantially for CoMFA datasets 4 and 5, reaching a final value of 0.771, which is indicative of a statistically strong correlation between the CoMFA energy field data and the biological activity of the compounds in the model. Nineteen of the 20 compounds of the series could be included in this final CoMFA model. Compound **13**, the only one in the series possessing a nitro-substituted phenyl ring, could not be satisfactorily incorporated in this CoMFA model. The conventional correlation coefficient r^2 of CoMFA dataset 5 was 0.988.

In order to confirm that this method of conformer selection was superior to a random selection, two CoMFA QSAR models were computed from randomly selected sets of one conformer per compound using the same compounds as in CoMFA dataset 5. The maximum crossvalidated r^2 values obtained in these two experiments (0.243 and 0.180) were markedly worse than the maximum crossvalidated r^2 of CoMFA dataset 5 (0.771), showing that the stepwise selection process did lead to better correlations.

If, alternatively, the selection was made of the minimum energy conformation of all compounds of CoMFA dataset 5, the correlation again deteriorated (maximum crossvalidated $r^2 = 0.619$). Clearly then, selection of the more stable conformation was better than a random selection but the iterative selection out-performed it. This result seems to be in natural agreement with the well-

TABLE 2
CoMFA MODELS DERIVED FROM STYRENES **1–20** (PROBE ATOM C sp^3-HYBRIDIZED, CHARGE +1.0)

CoMFA data set no.	Compounds	Number of compounds	Number of conformers	Max. crossval. r^2 (for number of components)	Relative contributions	
					Steric (%)	Electrostatic (%)
1	**2, 5, 19**	3	3	−0.534 (1)	74.5	25.5
2	**2, 5, 6, 12, 19**	5	5	−0.024 (2)	68.1	31.9
3	**2, 5, 6, 10, 12, 14, 15, 17–19**	10	10	0.178 (1)	79.0	21.0
4	**2, 3, 5–12, 14–20**	17	17	0.713 (6)	75.1	24.9
5	**1–12, 14–20**	19	19	0.771 (6)	75.2	24.8
6	**2, 5, 6, 10, 12, 14, 15, 17–19**	10	20	0.631 (5)	71.0	29.0
7	**2–12, 14–20**	18	30	0.813 (8)	70.3	29.7
8	**1–12, 14–20**	19	34	0.821 (7)	71.5	28.5

Data sets 1–5: single-conformer CoMFA models; data sets 6–8: multiple-conformer CoMFA models.

known fact (e.g. [44]) that the receptor–ligand interaction may lead to the ligand being bound in a conformation that is not its global energy minimum in the unbound state.

The choice of the initial CoMFA dataset and the threshold values for inclusion of conformers might seem to contain a certain degree of arbitrariness. However, this appeared to have very little influence on the final results. When variations of the iterative process were performed, e.g. by starting with a different initial dataset or using different threshold values, it turned out that a conformer not selected in a given step would usually be selected in one of the subsequent steps. Very similar final datasets were reached in all those cases, showing that the final CoMFA QSAR model is markedly robust against detail changes in the procedure. A noticeable independence of CoMFA of arbitrary choices of various kinds of parameters and strategies has been found in other studies (e.g. [29]).

1.2. Multiple-conformer CoMFA models

In order to investigate the influence on CoMFA of the (fixed) orientation of the ring hydroxyl hydrogens, which has been discussed in the Materials and Methods section, a second series of CoMFA models was built with C^+ as the probe atom. In this series, the initial CoMFA dataset (dataset 6 in Table 2) contained *two* conformers per molecule. Only the symmetrical molecules were used for this dataset, but both orientations of the ring hydroxyl hydrogens were included. This method effectively 'despecifies' the hydroxyl hydrogens' positions. It thereby tries to simulate a situation that was felt to be more physical than that of immobilized hydroxyl hydrogen positions at ambient temperatures. At the same time, it offers the advantage of circumventing the difficulties of the conformer selection process for the symmetrical molecules. For those symmetrical molecules (compounds **10, 14, 15, 17** and **18**), for which the rotation of the phenyl ring is not the only conformational degree of freedom, the two conformers with the lowest energy were selected. This was done in order not to bias the CoMFA process a priori by an unequal number of conformers for any of the symmetrical compounds, which might emphasize structural features of different compounds in an unequal way. It turned out that in each case this criterion led to the selection of two conformations differing only in the orientation of the phenyl ring and not in the vinyl moiety.

The same stepwise procedure as described earlier for the single-conformer CoMFA datasets (datasets 1–5 in Table 2) was utilized for deriving the final multiple-conformer CoMFA models (datasets 7 and 8 in Table 2). The selection of the conformers of the unsymmetrical molecules was based solely on the prediction accuracy for biological activity, and therefore, 1–4 conformers of those compounds were selected. This approach seems justified by the fact that no conformer can be ruled out a priori as a possible ligand to the receptor site, and it therefore seemed appropriate to include all conformers that had a predicted K_i of similar accuracy. For the first step, the threshold for inclusion of conformers was (arbitrarily) set to a factor of ≈ 2 between the prediction and the actual K_i ($\Delta pK_i \approx 0.3$). This was later relaxed to ≈ 5 ($\Delta pK_i \approx 0.7$) in order to allow more structures to be included. Nineteen out of the 20 compounds of the series (with a total of 34 conformers) could be included in this way in the final CoMFA dataset 8. Its maximum crossvalidated r^2 (0.821) clearly exceeded that of the corresponding final single-conformer CoMFA dataset 5 (0.771). The conventional r^2 of CoMFA dataset 8 was 0.981. It is worthwile noting that the maximum crossvalidated r^2 of the initial multiple-conformer CoMFA dataset 6 (0.631) is substantially higher than that of the corresponding initial single-conformer CoMFA dataset 3 (0.178).

TABLE 3
PREDICTION OF K_i VALUES FOR INDIVIDUAL CONFORMERS (C^+ PROBE ATOM)

Compound	K_i (μmol) predicted from								Measured K_i (μmol)
	Single-conformer CoMFA model (dataset 3 in Table 2)				Multiple-conformer CoMFA model (dataset 6 in Table 2)				
	for conformer				for conformer				
	A	B	C	D	A	B	C	D	
1	55.6	27.0	34.4	27.8	72.3	20.0	46.2	15.8	0.85
2	–[a]	X[b]			X	X			1
3	35.1	1.92			5.95	3.19			2
4	130	61.1	32.0	22.8	207	63.6	39.8	10.6	2.3
5	–	X			X	X			3.3
6	–	X			X	X			10
7	22.7	7.77			10.6	1.96			11
8	247	83.7	66.3	45.4	243	68.5	59.8	27.5	18
9	15.1	5.58	14.3	6.05	21.4	14.7	21.3	22.4	24
10	–	–	–	X	X	X	–	–	47
11	16.1	5.86			123	5.64			67
12	X	–			X	X			77
13	2.33	0.64			1.39	0.15			150
14	–	X	–	–	X	X	–	–	166
15	X	–	–	–	X	X	–	–	233
16	116	304	29.1	145	286	195	43.4	62.1	267
17	X	–	–	–	X	X	–	–	833
18	X	–	–		X	X	–		833
19	X	–			X	X			1333
20	514	81.5			2320	134			2500

[a] – = Conformer not used in building the CoMFA model.
[b] X = Conformer used in building the CoMFA model.

These results might, however, be affected by a possible bias of the crossvalidation procedure toward the multiple-conformer approach due to the fact that, after leaving out one structure, the other conformer(s) of this compound still remained in the CoMFA dataset, and therefore a more stringent test of the true predictive capability of CoMFA models for multi-conformer molecules was applied. In this case, the single-conformer dataset 3 and the multiple-conformer dataset 6 (both from Table 2) were compared as to their predictive accuracy (defined below) for the 10 unsymmetrical compounds, which were not part of either dataset. Table 3 shows the actual biological activities as well as the predicted activities of all conformers of those molecules derived from both CoMFA datasets. Inspection of the results in Table 3 reveals that predicted K_i values on the right half of the Table (multiple-conformer CoMFA model) are almost always more accurate than those on the left half (single-conformer CoMFA model), and some of these predictions (for compounds **7, 9, 16** and **20**) were in fact remarkably accurate (e.g. for compound **7**, the predicted K_i of conformer A was 10.6 μmol, while the measured K_i was 11 μmol). The average predictive accuracy, A_p, defined* as the multiplicative average of the relative accuracies of prediction for the most accurately predicted (m.a.p.) conformer of each compound,

$$A_p = \sqrt[n]{\prod_{j=1}^{n} q_j}, \quad q_j = \begin{cases} K_{i_{m.a.p.}}/K_{i_{act}}, & \text{if } K_{i_{m.a.p.}} \geq K_{i_{act}} \\ K_{i_{act}}/K_{i_{m.a.p.}}, & \text{if } K_{i_{act}} > K_{i_{m.a.p.}} \end{cases}$$

was calculated for both CoMFA datasets mentioned. (The lower this value, the higher the overall accuracy of the CoMFA model. The minimum value of 1.0 would be obtained in the case of perfect predictions for all molecules). A_p was 4.426 for the single-conformer CoMFA model (dataset 3), and 2.968 for the multiple-conformer CoMFA model (dataset 6). This clearly demonstrates the higher predictivity of the multiple-conformer CoMFA model. Comparing this approach to the averaging procedure mentioned in the Introduction [30], we note here that for the multiple-con-

*This value is somewhat similar to the PRESS value (PREdictive Sum of Squares) in CoMFA, since

$$\text{PRESS:} = \sum_{j-1}^{n} (pK_{i_{pred}} - pK_{i_{act}})_j^2$$

whereas

$$\ln A_p = \frac{1}{n} \sum_{j-1}^{n} |pK_{i_{m.a.p.}} - pK_{i_{act}}|_j$$

A_p, however, reflects better the *overall* predictive power of a CoMFA model, if one, for example, deems the case of a series of perfectly accurate predictions with the exception of one very inaccurate prediction better than a series of predictions that are all moderately inaccurate. To give an example, consider the case of nine novel compounds considered for synthesis, for which either

(a) eight K_i values are predicted perfectly accurately ($\Delta pK_i = 0$), and one prediction is inaccurate by a factor of 1000 ($\Delta pK_i = 3$); or

(b) all nine K_i values are predicted with an inaccuracy of 10 ($\Delta pK_i = 1$).

The PRESS is identical in both cases (a) and (b) (PRESS = 9), whereas the ln A_p values are significantly different [(a) ln A_p = 0.333 (A_p = 2.15); (b) ln A_p = 1 (A_p = 10)]. Obviously, based on (a), maximally one compound would falsely be excluded from (or included in) synthesis, whereas in case (b), depending on the individual values, this could happen for all compounds in the synthesis project. Thus A_p seems to us to better reflect the practical value of a CoMPA model in a real drug development situation.

former CoMFA model, the averaged K_i values were less accurate than the prediction for the most accurately predicted individual conformer in all cases, except one (**11**). The accuracy in ranking the compounds was also worse when average values were used.

2. O^- probe atom

The very low maximum crossvalidated r^2 of the initial CoMFA models calculated with the C^+ probe (datasets 1–3, Table 2) as well as the possibility that the electrostatic contribution was not handled optimally (suggested by the poor fit of the nitro-phenyl bearing compound **13** into the C^+ probe CoMFA models), indicated a potential benefit of investigating different probe atoms.

A probe atom was sought that was deemed more likely to resemble those found on a typical enzyme surface than C^+. An sp^3-hybridized oxygen atom bearing a charge of -0.4 (the charge of a hydroxyl oxygen as calculated by the Gasteiger algorithm in SYBYL) was investigated.

CoMFA analyses were repeated with this probe atom in much the same way as described for the C^+ probe, and the CoMFA models derived are shown in Table 4. It can be seen that the maximum crossvalidated r^2 of the single-conformer CoMFA datasets 1–5 (Table 4) reached higher values more quickly than those of the corresponding C^+ CoMFA datasets (Table 2). Further, this series more satisfactorily accounted for the nitro-phenyl bearing compound **13** than the C^+

TABLE 4
CoMFA MODELS DERIVED FROM STYRENES **1–20** (PROBE ATOM O sp^3-HYBRIDIZED CHARGE -0.4)

CoMFA data set no.	Compounds	Number of compounds	Number of conformers	Max. crossval. r^2 (for number of components)	Relative contributions	
					Steric (%)	Electrostatic (%)
1	**2, 6, 18, 19**	4	4	−0.020 (1)	87.9	12.1
2	**2, 5, 6, 10, 12, 17–19**	8	8	0.315 (3)	88.8	11.2
3	**2, 3, 5–7, 10–12, 17–20**	12	12	0.694 (3)	90.5	9.5
4	**2, 3, 5–12, 14–20**	17	17	0.747 (4)	90.3	9.7
5	**1–20**	20	20	0.743 (5)	92.0	8.0
6	**2, 5, 6, 10, 12, 14, 15, 17–19**	10	20	0.622 (8)	87.7	12.3
7	**2–12, 14–20**	18	29	0.810 (8)	89.3	10.7
8	**1–12, 14–20**	19	33	0.816 (10)	91.3	8.7

Data sets 1–5: single-conformer CoMFA models; data sets 6–8: multiple-conformer CoMFA models.

CoMFA datasets and **13** was therefore included in the final single-conformer O^- CoMFA dataset 5 (Table 4). The maximum crossvalidated r^2 of this dataset was 0.743. Its conventional r^2 was 0.944.

A series of multiple-conformer CoMFA models (datasets 6–8, Table 4) yielded similar results as the corresponding C^+ probe series. The maximum crossvalidated r^2 of the final multiple-conformer dataset 8 was 0.816. Its conventional r^2 was 0.981. It is noteworthy that for the oxygen probe, both the single-conformer and the multiple-conformer CoMFA datasets yielded the same selection of most accurately predicted conformers, whereas this was not the case for the carbon probe atom (see e.g. the predicted K_i values for compounds **9** and **16** in Table 3). The average predictive accuracy A_p was calculated here, too, for both a single-conformer dataset and a multiple-conformer dataset consisting only of symmetrical molecules. The A_p values were 5.034 and 2.832, respectively, demonstrating the higher predictivity of the multiple-conformer CoMFA models to an even stronger degree than had been the case with the C^+ probe.

The same two types of validation of this conformer selection method were carried out as had been done for the C^+ CoMFA models. Two different random selections yielded negative maximum crossvalidated r^2 values (-0.003 and -0.272). A CoMFA model built with the minimum energy conformers for all compounds yielded a maximum crossvalidated r^2 of 0.406, which is to be compared to the maximum crossvalidated r^2 of 0.743 of CoMFA dataset 5 (Table 4) derived by the iterative selection method. The ratio of these two numbers is considerably larger than the corresponding ratio in the C^+ CoMFA model series.

These results suggest that the oxygen probe, while not necessarily yielding higher final maximum crossvalidated r^2 values than the carbon probe, is more discriminating for the correct selection of conformers and might in this way be more advantageous for a stepwise selection process.

3. Neutral C probe atom

The results obtained using the two probe atoms C^+ and O^- seemed to allow the hypothesis that either a neutral or minimally charged probe atom might prove useful. Therefore, a multiple-conformer CoMFA model was constructed that comprised exactly the same 20 conformers of the 10 symmetrical molecules as the initial multiple-conformer CoMFA datasets (No. 6 in Tables 2 and 4) of both previous probe atoms. Here, however, an uncharged sp^3-hybridized carbon atom was used. The resulting maximum crossvalidated r^2 of 0.489 was considerably lower compared to the two former probe atoms, which gave 0.631 and 0.622, respectively (datasets 6 in Tables 2 and 4). The average predictive accuracy, A_p, was 4.912 (as compared to values of 2.968 and 2.832 for C^+ and O^-, respectively), indicating a much lower predictivity of this CoMFA model. These results suggest that a contribution from the electrostatic field of the molecules seems to be essential for a good CoMFA QSAR model, and therefore use of the neutral probe atom was not investigated further.

4. Prediction of K_i values and active conformations

4.1. General procedure

A CoMFA dataset with a high crossvalidated r^2 is suggestive of a statistically strong correlation between the 3D structure and the (known) biological activity of the compounds within this train-

ing set. However, this in itself does not provide a means for unambiguous prediction of the activity of conformationally flexible molecules that are not part of the training set. The problem is that different conformers of one molecule have to be entered as different molecules into CoMFA, and therefore multiple predictions will result for the same compound. The CoMFA methodology, as it stands, does not offer a direct means of determining *which one* of those conformations is responsible for the biological behaviour of the compound.

As has been described earlier, the approach of choosing between conformations on the basis of their respective potential energies failed, because such pre-selection eroded the strength of the evolving structure–activity relationship. External criteria being unavailable for the resolution of this problem, recourse was made instead to a solution based on internal consistency in the CoM-

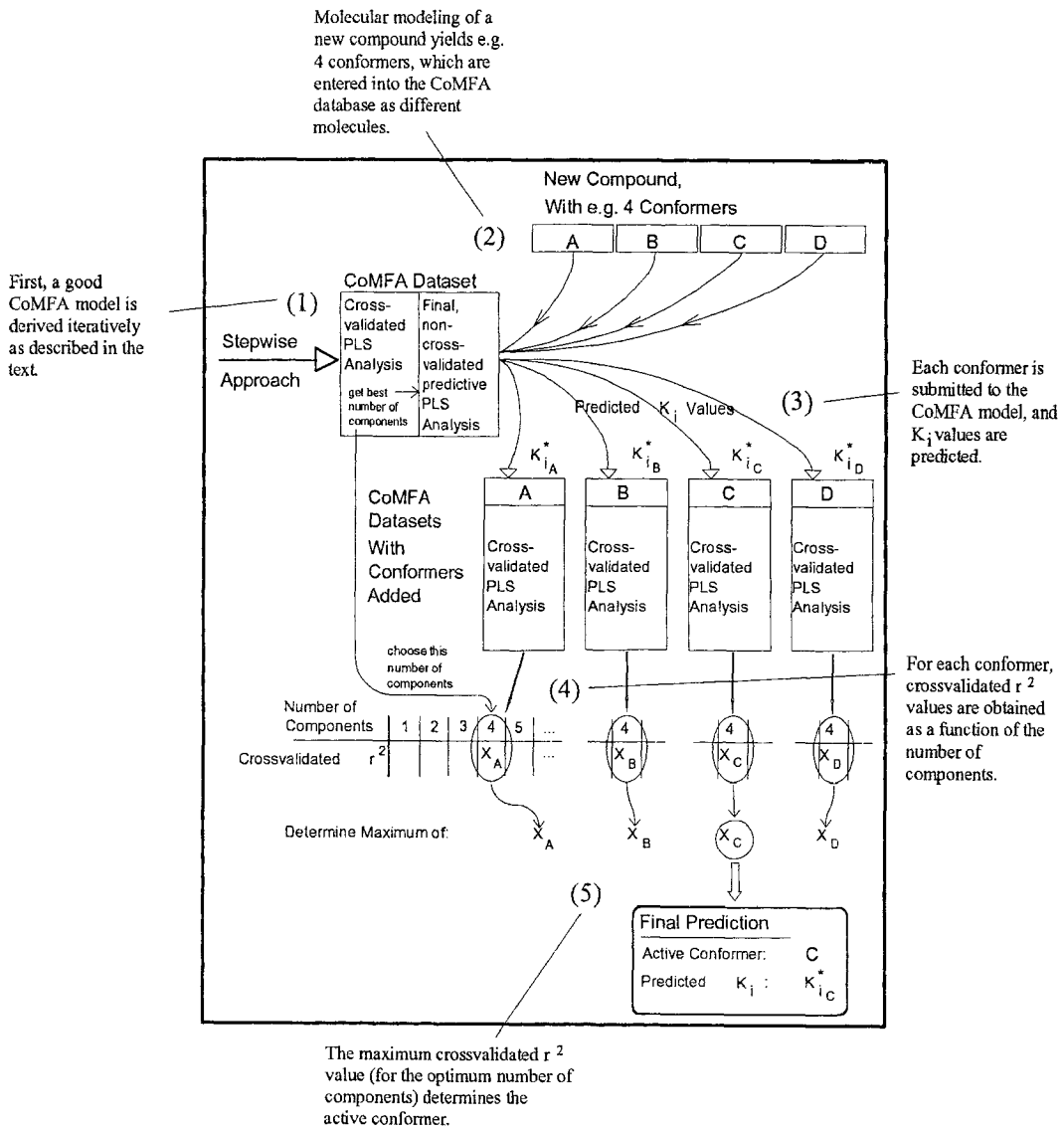


Fig. 2. Prediction of K_i and active conformation for new compound.

FA datasets. The basic idea was to calculate, using CoMFA, which one of the conformers of a new compound fits best into a pre-existing CoMFA model. For a compound having k conformers, for which a prediction of the active conformation and the K_i value is sought, this method can be outlined as follows (the arabic numerals in parentheses correspond to those in Fig. 2).

(i) As a preliminary step, a predictive CoMFA model with as high a predictivity as possible has to be derived (1). To achieve this, the iterative approach described in the preceding chapters may, for example, be employed, if it is applicable to the set of compounds under study. The number of components, for which this optimal model was derived, is recorded for future reference (step (vi)).

(ii) 3D atomic coordinates have to be derived for the new compound. This will usually be done using the molecular mechanics software, but other sources of 3D coordinates could also be used (such as X-ray crystallography or ab initio calculations). Since the premise was made that the new compound possesses conformational flexibility, a set of different conformers has to be derived that would appear most likely to contain the actual bio-active conformer. If the 3D coordinates have been generated via molecular modeling, this would usually include all local energy minimum structures within a certain energy range from the global energy minimum (2). For 3D coordinates obtained from other sources, different criteria might be employed. We assume that a set of k different conformers has been derived for the new compound.

(iii) For each conformer of the new compound, a prediction of its K_i value is derived from the predictive CoMFA model (3).

(iv) Each conformer is then added separately to the original CoMFA dataset of structures. This yields k new datasets, each one of those having been augmented by a different structure (i.e. conformer of the new compound). The predicted K_i values (step (iii)) are entered in the CoMFA data tables as the biological activity data of the new compound's conformers.

(v) k new CoMFA models are built with the k new datasets. Only the crossvalidation CoMFA is performed in each case (4).

(vi) The crossvalidated r^2 values at the number of components used to build the original CoMFA model (step (i)) are compared to each other (5). The conformer that yields the maximum value is assumed to be the active conformer, and its predicted K_i value (step (iii)) is taken as the new compound's predicted biological activity.

4.2. Test of the method with the given dataset

In order to allow a test of this method with the styrene series (**1–20**), a slight modification of this technique had to be employed. Since no biological data were available for molecules other than **1–20** in the assay system used [32], no test with a truly new compound could be performed, and so the unsymmetrical molecules were used as test cases.

As the starting point, the final CoMFA dataset of the multiple-conformer series for the O^- probe (i.e. No. 8 in Table 4), which contains 19 compounds, was selected. Each one of the eight compounds listed in Table 5 was taken out of the data set in turn, and then treated according to the procedure just described as if it were a new compound. (This means, of course, that eight, slightly different, initial predictive CoMFA models had to be constructed for each of the eight different reduced sets of compounds. This, however, can only diminish the internal consistency of this approach, and will certainly not produce artificially good results). The number of compo-

TABLE 5
FINAL PREDICTION OF K_i VALUES

Compound	Predicted			Measured		
	K_i (μmol)	pK_i	Rank	K_i (μmol)	pK_i	Rank
3	3.05	5.516	2	2	5.699	1
4	0.38	6.420	1	2.3	5.638	2
7	6.60	5.180	3	11	4.959	3
8	6.78	5.169	4	18	4.745	4
9	8.90	5.051	5	24	4.620	5
11	38.0	4.420	6	67	4.174	6
16	368	3.434	7	167	3.573	7
20	2409	2.618	8	2500	2.602	8

nents used in steps (i) and (vi) of these analyses seems to be the single most critical factor in the prediction process described here. As mentioned earlier, this number affects both crossvalidated r^2 as well as the predictive PLS (non-crossvalidated) analyses, but in different ways. In this study, the maximum crossvalidated r^2 values were typically obtained for a number of components falling in a range between 7 and 10, but the best predictive PLS results were usually found when four components were used (the value where the curve of the crossvalidated r^2 as a function of the number of components started to level off after its steep initial increase; see also the footnote * on page 492). This value was therefore used for building all (predictive) CoMFA models in this analysis. Comparison of the *maximum* crossvalidated r^2 values failed to produce a correct ranking of the test compounds. In contrast, comparison of the crossvalidated r^2 values for four components led to a very accurate prediction of the K_i values and an almost perfect ranking, as shown in Table 5. The average predictive accuracy, A_p, was 2.02, and the only deviation in the ranking was the swapping of the compounds with the two lowest (and nearly identical) measured K_i values.

DISCUSSION

The general idea of the approach that has been described in this paper is to use a subset of compounds to predict activities of multiple conformers of other compounds and then include the most accurately predicted conformer(s) in the CoMFA dataset in the next step of an iterative procedure. A set of molecules which possess different degrees of internal symmetry, as has been used in this study, naturally lends itself to this approach. This criterion is, however, no prerequisite, and other criteria for subdividing a set of compounds can easily be conceived, such as the availability of 3D coordinates from X-ray crystallography.

If no such experimental structural data are available, however, molecular models derived from some kind of computation have to be used. It is obvious that the quality of any CoMFA study performed on such models cannot be better than the quality of those models themselves. Since in general there is no simple way of determining the absolute quality of molecular models*, one

*Unless 3D structures determined by X-ray crystallography are available, and even those are not of unlimited validity for the molecular shape in the biological environment (e.g. [44]).

usually has to use the 3D structures such as they are produced by the conformational search and minimization procedures performed by the molecular modeling software. However, the CoMFA technique itself offers a certain degree of quality assessment of the molecular models used. It has been shown [42] that CoMFA is a 'conservative' technique that is very unlikely to pick up a chance correlation in a given dataset (yielding a false-positive answer). A high correlation (high cross-validated r^2) detected by CoMFA between the measured biological activity and the molecular fields therefore strongly suggests that at least reasonably accurate molecular models have been used.

Although the choice of the structures for the small initial CoMFA dataset seems to suffer from arbitrariness, the mentioned relative robustness of the final CoMFA QSAR model against this very choice demonstrates the 'objectivity' of this self-consistency type of method in CoMFA. Having the program calculate the best fit of conformers into the CoMFA model seems to be less biased than e.g. simply selecting the lowest energy structure (this being a calculated structure itself in most instances). The crossvalidated r^2 values that were obtained when comparing those two approaches support this view. For the oxygen probe atom e.g., a maximum crossvalidated r^2 of 0.743 was obtained for the single-conformer CoMFA model for which each compound's conformer had been selected with the iterative technique. The corresponding CoMFA model built with the minimum energy conformers of all compounds yielded a maximum crossvalidated r^2 of only 0.406.

The prediction method for novel compounds that has been described seems to offer as objective a procedure as possible, within the CoMFA methodology, for picking the biologically active conformation of a flexible molecule. Due to its principle of adding a single structure (conformer) at a time to a pre-existing dataset, it clearly finds its limitation in the size of this dataset. It is likely that if one has a large dataset to begin with, then the addition of one more compound will only yield a small perturbation of the model, in which case it may become difficult to reliably select the most internally consistent conformer. This point suggests that it may be advisable not to artificially inflate the pre-existing dataset in order to ensure a sufficient sensitivity of the model to the addition of the different conformers.

The points discussed so far lead one naturally to reflect on the possible benefits of tailoring a CoMFA search for a new drug specifically for the approach described in this paper. This would particularly entail selection of the standard compounds according to the criteria of (1) symmetries found in the molecules, and (2) the different possible substituent positions in the class of the molecules considered for synthesis being systematically covered by substituents of interest. This is important since the prediction of the biological activity of a new compound with substituents in areas not covered by the standard compounds ('predicting something from nothing') is likely to be unreliable. Results of the application of the methodology described in this publication in the development of new drugs, especially with PTK inhibitory activity, will be reported in future papers.

CONCLUSIONS

A highly correlative QSAR model has been derived for the set of PTK substrate inhibiting compounds. It was shown that the stepwise approach to a CoMFA QSAR model is capable of dealing with conformationally ambiguous molecules. If the training set consists of molecules with a diffe-

rent number of conformational degrees of freedom, a successful approach seems to be the use of more symmetrical molecules to build initial CoMFA tables, and then to use these to select the active conformers of the less symmetrical molecules. Very low crossvalidated r^2 values in the first steps do not appear to be a cause of concern, since the final CoMFA tables exhibited high crossvalidated r^2.

Despecifying hydrogen atom positions by entering multiple conformers of molecules in the CoMFA datasets yielded final CoMFA models with higher predictive power than single-conformer datasets.

A probe atom of relatively small charge, sp^3-hybridized oxygen with a charge of -0.4, seems to exhibit higher selectivity for conformers than a sp^3-hybridized carbon probe atom with a charge of $+1.0$. However, some charge seems to be required, since a neutral sp^3-hybridized carbon probe atom yielded poorer results than either charged probe atom.

A technique has been devised that appears to be capable of deriving accurate predictions for conformationally flexible molecules and correct prospective ranking with respect to activity within a group of compounds of this type. To do this, separate CoMFA models were constructed for each conformer of a molecule whose activity is to be predicted. The crossvalidated r^2 of these CoMFA models were compared to each other, the maximum indicating the active conformer. It seems crucial to perform this comparison at exactly that number of components that was used in building the final, predictive CoMFA model. This method reproduced the ranking among our series of test molecules almost perfectly.

REFERENCES

1 Roberts, T.M., Kaplan, D., Morgan, W., Keller, T., Mamon, H., Piwnica-Worms, H., Druker, B., Cohen, B., Schaffhausen, B., Whitman, M., Cantley, L., Rapp, U. and Morrison, D., In Cold Spring Harbor Symposium on Quantitative Biology, Vol. LIII, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1988, pp. 161–171.

2 Ullrich, A. and Schlessinger, J., Cell, 61 (1990) 203.

3 Heldin, C.H., Betsholtz, C., Claesson-Welsh, L. and Westermark, B., Biochim. Biophys. Acta, 907 (1987) 219.

4 Anderson, G., Mihich, E., Nishimura, S., Terada, M. and Sugimura, T., Cancer Res., 49 (1989) 6852.

5 Felice, G., Horvath, A.R. and Kellie, S., Biochem. Soc. Trans., 18 (1990) 69.

6 Neal, D.E., Sharples, L., Smith, K., Fennelly, J., Hall, R.R. and Harris, A.L., Cancer, 65 (1990) 1619.

7 Toi, M., Nakamura, T., Mukaida, H., Wada, T., Osaki, A., Yamada, H., Toge, T., Niimoto, M. and Hattori, T., Cancer, 65 (1990) 1980.

8 Uehara, Y., Hori, M., Takeuchi, T. and Umezawa, H., Mol. Cell. Biol., 6 (1986) 2198.

9 Imoto, M., Umezawa, K., Komuro, K., Sawa, T., Takeuchi, T. and Umezawa, H., Jpn. J. Cancer Res., 78 (1987) 329.

10 Shiraishi, T., Domoto, T., Imai, N., Shimada, Y. and Watanabe, K., Biochem. Biophys. Res. Commun., 147 (1987) 322.

11 Kruse, C.H., Holden, K.G., Offen, P.H., Pritchard, M.L., Feild, J.A., Rieman, D.J., Bender, P.E., Ferguson, B., Greig, R.G. and Poste, G., J. Med. Chem., 31 (1988) 1768.

12 Gazit, A., Yaish, P., Gilon, C. and Levitzki, A., J. Med. Chem., 32 (1989) 2344.

13 Saperstein, R., Vicario, P.P., Strout, H.V., Brady, E., Slater, E.E., Greenlee, W.J., Ondeyka, D.L., Patchett, A.A. and Hangauer, D.G., Biochemistry, 28 (1989) 5694.

14 Isshiki, K., Imoto, M., Sawa, T., Umezawa, K., Takeuchi, T., Umezawa, H., Tsuchida, T., Yoshioka, T. and Tatsuta, K., J. Antibiot. (Tokyo), 40 (1987) 1209.

15 Onoda, T., Inuma, H., Sasaki, Y., Hamada, M., Isshiki, K., Naganawa, H. and Takeguchi, T., J. Nat. Prod. (Lloydia), 52 (1989) 1252.

16 Shiraishi, T., Kameyama, K., Imai, N., Domoto, T., Katsumi, I. and Watanabe, K., Chem. Pharm. Bull., 36 (1988) 974.

17 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., In Fauchère, J.L. (Ed.), QSAR: Quantitative Structure–Activity Relationships in Drug Design (Proceedings of the 7th European Symposium on QSAR) (Progress in Clinical and Biological Research, Vol. 291), Alan R. Liss, Inc., New York, NY, 1989, pp. 161–165.
18 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
19 Kellogg, G.E., Semus, S.F. and Abraham, D.J., J. Comput.-Aided Mol. Design, 5 (1991) 545.
20 Marshall, G.R., Mayer, D., Naylor, C.B., Hodgkin, E.E. and Cramer III, R.D., In Fauchère, J.L. (Ed.), QSAR: Quantitative Structure–Activity Relationships in Drug Design (Proceedings of the 7th European Symposium on QSAR) (Progress in Clinical and Biological Research, Vol. 291), Alan R. Liss, Inc., New York, NY, 1989, pp. 287–295.
21 Unwalla, R.J., Young, S.S. and Profeta Jr., S., In 199th ACS Meeting, Boston, MA, April 1990, Abstr. MEDI 34.
22 Allen, S.M., Tan, Y.-C., Trudell, M.L., Narayanan, K., Schindler, L.R., Martin, M.J., Schultz, C., Hagen, T.J., Koehler, K.F., Codding, P.W., Skolnick, P. and Cook, J.M., J. Med. Chem., 33 (1990) 2343.
23 Diana, G.D., Treasurywala, A.M. and Bailey, T.R., In 201th ACS Meeting, Atlanta, GA, April 1991, Abstr. MEDI 43.
24 Diana, G.D., Kowalczyk, P., Treasurywala, A.M., Oglesby, R.C., Pevear, D.C. and Dutko, F.J., J. Med. Chem., 35 (1992) 1002.
25 Carroll, F.I., Gao, Y., Rahman, M.A., Abraham, P., Parham, K., Lewin, A.H., Boja, J.W. and Kuhar, M.J., J. Med. Chem., 34 (1991) 2719.
26 Maret, G., El Tayar, N., Carrupt, P.-A., Testa, B., Jenner, P. and Baird, M., Biochem. Pharmacol., 40 (1990) 783.
27 Björkroth, J.-P., Pakkanen, T.A. and Lindroos, J., J. Med. Chem., 34 (1991) 2338.
28 Kim, K.H. and Martin, Y.C., J. Org. Chem., 56 (1991) 2723.
29 Greco, G., Novellino, E., Silipo, C. and Vittoria, A., Quant. Struct.–Act. Relat., 10 (1991) 289.
30 Kim, H.K. and Martin, Y.C., J. Med. Chem., 34 (1991) 2056.
31 SYBYL Molecular Modeling Software Version 5.2 Theory Manual, Tripos Associates, Inc., St. Louis, MO 63144, 1989, p. 1160.
32 Yaish, P., Gazit, A., Gilon, C. and Levitzki, A., Science, 242 (1988) 933.
33 Tripos Associates, Inc., 1699 South Hanley Road, St. Louis, MO 63144.
34 Clark, M., Cramer III, R.D. and Van Opdenbosch, N., J. Comput. Chem., 10 (1989) 982.
35 Gasteiger, J. and Marsili, M., Tetrahedron, 36 (1980) 3219.
36 Marsili, M. and Gasteiger, J., Croat. Chim. Acta, 53 (1980) 601.
37 Broyden, C.G., Fletcher, Y., Goldfarb, D. and Shanno, D.F., In Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (Eds.), Numerical Recipes: The Art of Scientific Computing, Cambridge University Press, Cambridge, 1986, pp. 301–307.
38 Nagy, P.I., Dunn III, W.J., Alagona, G. and Ghio, C., J. Am. Chem. Soc., 113 (1991) 6719.
39 Wold, S., Albano, C., Dunn III, W.J., Edlund, U., Esbenson, K., Geladi, P., Hellberg, S., Johannson, E., Lindberg, W. and Sjöström, M., In Kowalski, B. (Ed.), Chemometrics: Mathematics and Statistics in Chemistry, Reidel, Dordrecht, The Netherlands, 1984, pp. 17–95.
40 Wold, S., Ruhe, A., Wold, H. and Dunn III, W.J., SIAM J. Sci. Stat. Comput., 5 (1984) 735.
41 Stahle, L. and Wold, S., Prog. Med. Chem., 25 (1988) 292.
42 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., Quant. Struct.–Act. Relat., 7 (1988) 18.
43 SYBYL Molecular Modeling Software Version 5.2 Theory Manual, Tripos Associates, Inc., St. Louis, MO 63144, 1989, pp. 1176–1182.
44 Kubinyi, H., J. Cancer Res. Clin. Oncol., 116 (1990) 529.