Eur. J. Med. Chem. 35 (2000) 123−136

S0223523400001082/FLA

Original article

Validation of protein-based alignment in 3D quantitative structure–activity relationships with CoMFA models

Alexander Golbraikh, Philippe Bernard, Jacques R. Chrétien*

Laboratory of Chemometrics and Bioinformatics, University of Orléans, BP 6759, 45067 Orléans, France

Received 5 October 1999; accepted 9 November 1999

Abstract – The predictive capabilities of protein-based alignment (PBA) and structure-based alignment (SBA) comparative molecular field analysis (CoMFA) models have been compared. 3D quantitative structure–activity relationship (3D QSAR) models have been derived for a series of N-benzylpiperidine derivatives which are potent acetylcholinesterase (AChE) inhibitors interesting for Alzheimer’s disease. To establish a comparison with the classical SBA procedure, different assay models were derived by superposing ligand conformers that are docked to the AChE active site and by using the most active compound as the reference one. A Kohonen self organizing map (SOM) was applied to analyse the molecular diversity of the test set relative to that of the training set, in order to explain the influence of molecular diversity on the predictive power of the considered models. SBA 3D QSAR models have to be used to predict the inhibitory activity only for compounds belonging to subgroups included in the training set. The PBA 3D QSAR models appeared to have a higher predictability, even for compounds with a molecular diversity greater than that of the training set. This results from the fact that the protein helps to automatically select the active conformation which is fitting the 3D QSAR model. © 2000 Éditions scientifiques et médicales Elsevier SAS

CoMFA / 3D QSAR / Alzheimer’s disease / protein-based alignment / structure-based alignment / acetylcholinesterase

1. Introduction

Comparative molecular field analysis (CoMFA) is a 3D quantitative structure–activity relationship (3D QSAR) approach superior to most other QSAR methods as regards to their predictive capabilities [1] and was proposed in the middle of the 1970s by U. Burkert and N. Allinger [2]. Wide use of the method started with the contributions of R.D. Cramer III and G.R. Marshall et al. (see, for instance, [3, 4]) 12 years later. The idea underlying CoMFA is that differences in the molecular properties of compounds can often be explained by differences in the non-covalent fields surrounding the molecules. CoMFA is based on the Lennard-Jones steric and the Coulombic electrostatic field values, computed at the intersections of a lattice within a 3D region surrounding the molecules. Thus, each CoMFA descriptor is represented by steric or electrostatic field values at a certain grid point. These descriptors serve as independent variables in QSAR analysis. A regression-like technique, the partial least squares (PLS) method, is recommended [5] for deriving a 3D QSAR/CoMFA model to predict the biological activity. PLS was developed by H. Wold in the middle of the 1970s [6, 7], to handle two problems: (i) when the number of variables exceeds the number of samples many times and (ii) when usual regression methods, such as multiple linear regression (MLR), are not suitable according to the linear independence of the variables [5]. Since then, it has become a more and more popular statistical method successfully used in chemometrics [8, 9]. To test the robustness of the model, a cross-validation procedure is now required [5, 10]. Cross-validation consists of dividing a training set into a given number (*n*) of groups of compounds, approximately equal in size, and performing a PLS (or some other QSAR procedure) *n* times, but leaving out one of these groups each time to perform the calculations. Statistical parameters derived from this procedure are used to assess the quality of the model [11]. A more rigorous validation of the model consists of applying it to property prediction on compounds not represented in the training set [5, 10].

*Correspondence and reprints:
Jacques.Chretien@univ-orleans.fr

The first step of a CoMFA procedure is to obtain the compound active conformations responsible for the bioactivity and to align these conformations in space, in accordance with a postulated pharmacophore model, docking results or crystallographic data, etc. The decision remains as to which alignment is to be considered to perform the CoMFA procedure.

A structure-based alignment (SBA) deals with a set of ligands which are superimposed onto a so called reference molecule. It is usually the most rigid one, possessing a high affinity to the receptor, according to a proposed pharmacophore model. SBA of a series of ligands is a common approach for obtaining a mutual spatial arrangement of the pharmacophore elements when the 3D structure of the receptor is not known, or not used directly for the alignment. In some cases, especially if all compounds are flexible, this approach can lead to doubtful results owing to the fact that the chosen conformations remain far from the biologically active ones.

On the other hand, a protein-based alignment (PBA) involves a set of ligands docked to the active site of the protein, which optimizes the choice of the biologically active conformations, before superimposing them to each other according to their relative positions in the active site. Protein alignment is based on the known 3D structure of the receptor active site. If the 3D structure of a ligand bound to the protein is known, conclusions about the positions of ligand functional groups can be drawn and used for enlightening docking possibilities of other compounds of interest.

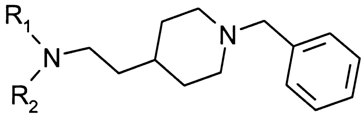
Nowadays, as the number of available protein 3D structures is increasing rapidly, the number of papers involved in docking procedures is growing similarly. For example, in the ‘PubMedline’ database, the keyword ‘Docking’ leads to 15 papers up to 1988, 150 up to 1994, and 313 up to 1997. The publication number has increased more than 20-fold over the past 10 years. A similar trend can be observed for publications with the keywords ‘CoMFA’ and ‘3D QSAR’, even if their number remains lower than that with the keyword ‘Docking’. PBA CoMFA is a new approach to 3D QSAR analysis, the total number of papers published up to now and registered in ‘MedLine’ with the keyword ‘CoMFA’ equals 149, and only 6 have both keywords ‘CoMFA’ and ‘Docking’ and they appeared only in the past four years. Furthermore, even if both keywords appear simultaneously in the same paper it does not imply that a PBA CoMFA procedure was applied.

Mouse AChE appeared to be an ideal candidate for applying automated docking to a series of N-benzylpiperidines, reversible AChE inhibitors, and obtaining their protein-based alignment. At the same time, AChE inhibitors play a crucial role in the contemporary approaches to the treatment of Alzheimer’s disease, namely senile dementia caused by a decrease in acetylcholine production according to the cholinergic hypothesis (see, for instance, [12, 13]). So the development of new models based on docking for the prediction of the AChE inhibitory activities could lead to a better understanding of the molecular mechanisms underlying their inhibitory activities, and as a result, to the development of new biologically active compounds as potential drugs.

This approach was proposed as an ‘ideal solution that would create an incontestable natural alignment’ [14]. Similar approaches, based on an anchor or base fragment, were recently developed for the softwares FlexX and AutoDock for automated docking [15, 16]. The leading fragment approach is also used in the Hammerhead software [17]. In contrast to these algorithms, in [14] only one atom, namely the quaternary nitrogen of the piperidine moiety, was used as an anchor. The position of this atom was suggested from crystallographic data [18]. A 3D QSAR model was derived, within the AChE active site, with help of the protein-based alignment of a training set involving 82 N-benzylpiperidines possessing a benzoyl or a phthalimide group, or with other substituents for these groups as well as the benzyl one (*tables I–III*). This model was validated with another sub-series of 29 N-benzylpiperidines. But in fact, most of them were N-benzylpiperidine-benzisoxazole derivatives not represented in the training set (*table IV*). Nevertheless the relative inhibitory activities were correctly predicted for each of the 29 compounds and prove the consistency of the model and the particular interest of the 3D QSAR/PBA CoMFA approach.

The aim of this paper was to validate this PBA CoMFA procedure. Then the topic was to estimate more deeply the difference between the CoMFA models based on protein-based alignments and those derived from structure-based alignment, i.e. the PBA CoMFA and the SBA CoMFA procedures. That is to say, is it possible to construct an SBA model that would have a predictive power comparable to the model derived from the protein-based alignment? [14]. Or in other words, is it possible to demonstrate that PBA CoMFA models, when accessible, have a better predictive power than the corresponding SBA CoMFA model? We started directly from the conformations obtained by the above mentioned automated docking procedure. The problem concerning how to obtain these conformations in some other way was not considered here. We already had the right ‘biologically active conformations’ and the only task was to try to superimpose them in a correct manner to obtain a good

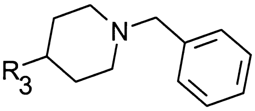
Table I. In vitro inhibition of AChE by *N*-benzylpiperidine derivatives.



Compound	R_1	R_2	IC_{50} (nM)	log (1/IC_{50}) (μM)
1	PhCO	H	560.0	0.25
2	*o*-CH_3 PhCO	H	1 000.0	0.00
3	*m*-CH_3 PhCO	H	470.0	0.33
4	*p*-CH_3 PhCO	H	180.0	0.74
5	*o*-NO_2 PhCO	H	880.0	0.06
6	*m*-NO_2 PhCO	H	230.0	0.64
7	*p*-NO_2 PhCO	H	55.0	1.26
8	*p*-OCH_3 PhCO	H	88.0	1.06
9	*p*-CHO PhCO	H	120.0	0.92
10	*p*-Cl PhCO	H	180.0	0.74
11	*p*-F PhCO	H	85.0	1.07
12	*p*-CH_3CO PhCO	H	51.0	1.29
13	*p*-($PhCH_2SO_2$) PhCO	H	29.0	1.54
14	*o*-pyridineCO	H	800.0	0.10
15	*m*-pyridineCO	H	69.0	1.16
16	*p*-pyridineCO	H	39.0	1.41
17	$C_6H_{11}CO$	H	1 600.0	–0.20
18	$PhCH_2$	H	46 000.0	–1.66
19	PhCO	CH_3	170.0	0.77
20	PhCO	C_2H_5	130.0	0.89
21	PhCO	$PhCH_2$	940.0	0.03
22	PhCO	Ph	35.0	1.46
23	*p*-($PhCH_2SO_2$) PhCO	CH_3	0.6	3.22
24	*p*-($PhCH_2SO_2$) PhCO	C_2H_5	0.3	3.52
25	*p*-($PhCH_2SO_2$) PhCO	Ph	0.6	3.22
26	*p*-OCH_3 PhCO	Ph	590.0	0.23
27	*p*-F PhCO	Ph	18.0	1.74
28	*p*-NO_2 PhCO	Ph	5.4	2.27
29	*p*-pyridineCO	Ph	64.0	1.19
30	$C_6H_{11}CO$	Ph	9 400.0	–0.97
31	CH_3CO	Ph	52.0	1.28
32	CH_3CH_2CO	Ph	830.0	0.08
33	CH_3CO	*m*-OCH_3 Ph	46.0	1.34
34	CH_3CO	*p*-OCH_3 Ph	700.0	0.15
35	CH_3CO	*m*-F Ph	65.0	1.19
36	CH_3CO	*p*-F Ph	205.0	0.69
37	CH_3CH_2	Ph	12 000.0	–1.08
38	CH_3CO	*p*-pyridine	108.0	0.97
39	CH_3CO	CH_3	660.0	0.18

CoMFA model as regards its predictive capabilities. Furthermore, this paper addresses experimental design [19] in computer aided molecular design problems for the choice of an appropriate test set of compounds to validate the model, i.e. to demonstrate its robustness on a rational basis.

Table II. In vitro inhibition of AChE by *N*-benzylpiperidine derivatives.



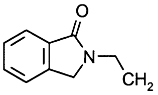
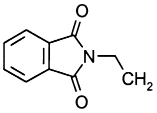
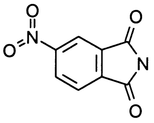
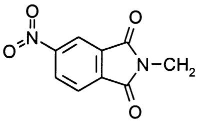
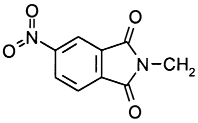
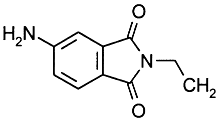
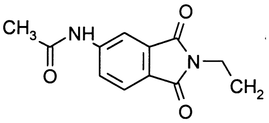
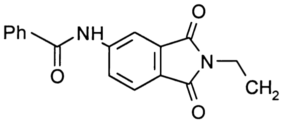
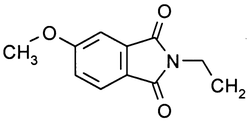
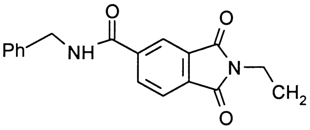
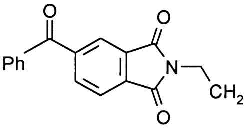
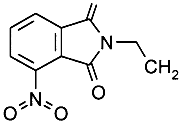
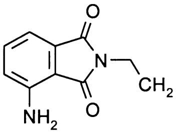
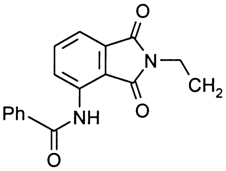
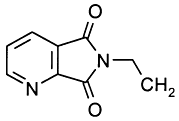
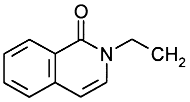
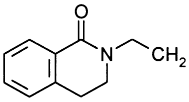
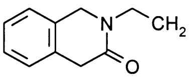
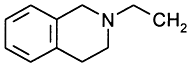
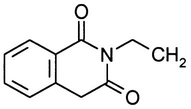
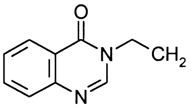
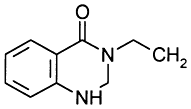
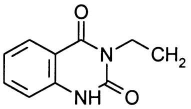
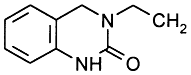
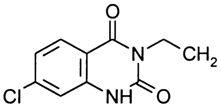
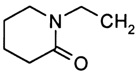
Compound	R_3	IC_{50} (nM)	log $(1/IC_{50})$ (μM)
40	$PhCO(CH_2)_3$	530.0	0.28
41		98.0	1.01
42		30.0	1.52
43		27 000.0	–1.43
44		3 000.0	–0.48
45		12.5	1.90
46		8.8	2.06
47		2.8	2.55
48		1.2	2.92
49		8.0	2.10
50		2.2	2.66
51		2.4	2.62
52		9.0	2.05
53		11.0	1.96
54		340.0	0.47
55		13.0	1.89
56		1 100.0	–0.04
57		1 000.0	0.00
58		17.0	1.77

Table II. Continued.

59		1 600.0	–0.20
60		23.0	1.64
61		1 200.0	–0.08
62		800.0	0.10
63		4.2	2.38
64		13.0	1.89
65		4.5	2.35
66		270.0	0.57

2. Methods

2.1. Software

All 3D QSAR models or assay models presented in this study were derived using the SYBYL package from Tripos [1]. The minimization of the conformations was performed using Gasteiger-Hückel atomic charges.

The electrostatic CoMFA fields were calculated using MOPAC AM1 charges. The CoMFA region was set to include all the molecules with margins of 3.0–4.0 (and was similar to the one defined in [14], the grid step being equal to 1 Å. The PLS method [8] was used for constructing the possible CoMFA models, the log(1/IC_{50}) values of the compounds were used as dependent variable values. Here the IC_{50} values express the inhibitory activities of the compounds. A leave-one-out cross-validation procedure was used to obtain the optimal number of PLS components. The models were then used to predict the inhibitory activities of the test set.

Compound **24** (*table I*) is the most active of the AChE inhibitors considered previously and in this study [14]. Hence, its docked to AChE conformation was used as a reference in all superpositions. The pharmacophore model consisted of a quaternary piperidine nitrogen, an amide group and a benzene ring. It was selected in accordance with the findings by Sugimoto et al. [20, 21] that both the benzoyl (or phthalimide) and the N-benzylpiperidine moieties are important for binding to AChE and inhibiting it. Three different types of superposition A, B and C, depending of the considered pharmacophore elements, were envisaged to perform the structural alignments (*figure 1*).

2.2. Compound alignment A

The three elements of the pharmacophore model were used. The two carbonyl atoms of the amide were included in the model in order to take into account the possibility of a hydrogen bond forming between a carbonyl oxygen and some of the protein hydrogens. For the compounds from *table II* possessing two phtalimide carbonyl groups, only the one closest to that of molecule **24** was taken into account in the PBA case. For compounds **56**, **57** and **59** with only one carbonyl group, only the corresponding carbon atoms were taken into consideration. For compounds **76**, **77**, **79** and **82** (*table III*) without a benzene ring, none of the corresponding atoms were taken into account. As far as the test set presented *table IV* is concerned, instead of the corresponding benzisoxazole oxygen, the following atoms were taken for structural alignment: (i) the corresponding sulfur atom of compound **21**, (ii) the sp^2 carbon atom closest to the imine group of compound **22**, and (iii) the corresponding nitrogens for compounds **23** and **24**.

2.3. Compound alignment B

Superposition was performed taking into account the quaternary piperidine nitrogen and the amide group, according to the rules formulated above as to these pharmacophore elements.

2.4. Compound alignment C

It was noticed that in the protein-based alignment the positions of the quaternary nitrogen and benzene ring are similar for all compounds, except for those from *table III*. So only these two pharmacophore elements were used for superposition.

2.5. Different calculations

Two kinds of calculations were performed. First, SBA 3D QSAR models were constructed to evaluate their predictive capabilities for the compounds belonging to the same group (or sub-series) of compounds, as in the training set. Three different models were constructed to this aim. In all these models, only the docked to AChE conformations of the compounds were used. The numbers at the beginning of the paragraphs below will be referred as the numbers of the assay models.

– Assay model 1

Every fifth compound (i.e. with numbers **5**, **10**, **15**, etc, see *tables I–III*) was excluded from the initial training set of the 82 benzyl-piperidines [14] to create a new test set of 16 compounds, selected at random. Thus, the new training set contained 66 compounds. All the compounds were superimposed according to alignment A (*figure 1*). A 3D QSAR model was then constructed using compounds belonging to the training set and tested on the remaining 16 compounds. The quaternary nitrogens were protonated for all compounds.

– Assay model 2.

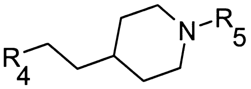
Similar calculations were performed for the 66 compounds from *tables I* and *II*. The compounds were superimposed according to alignment C (*figure 1*), and every fifth compound was then excluded from the training set to form a new test set consisting of 13 compounds. The quaternary nitrogens were protonated.

– Assay model 3

The same calculations were performed with the protonated pyridine nitrogens of compounds **14–16**, **29** and **38** from the training set (*table I*).

The second part of this study was devoted to the assessment of 3D QSAR models based on structural alignment as regards to their predictive capabilities for compounds belonging to subgroups not represented in the training set. Recently, it was shown that the protein-based QSAR model constructed for 82 N-benzylpiperidine

Table III. In vitro inhibition of AChE by N-piperidine derivatives.



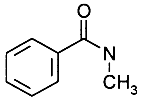
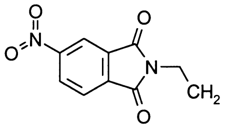
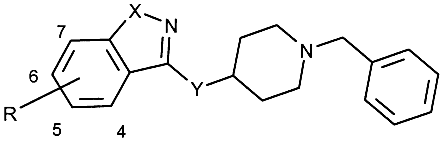
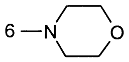
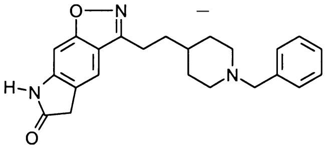
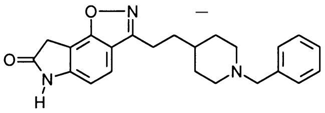
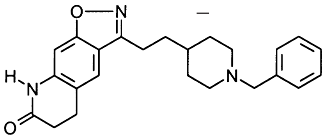
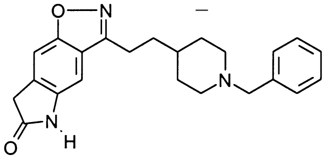
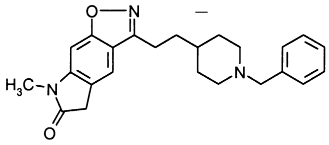
Compound	R_4	R_5	IC_{50} (nM)	log (1/IC_{50}) (μM)
67		*o*-Ch$_3$ Bzl	770.0	0.11
68	id	*m*-Ch$_3$ Bzl	145.0	0.84
69	id	*p*-Ch$_3$ Bzl	41 000.0	–1.61
70	id	*o*-NO$_2$ Bzl	14 000.0	–1.15
71	id	*m*-NO$_2$ Bzl	370.0	0.43
72	id	*p*-NO$_2$ Bzl	3 300.0	–0.52
73	id	PhCH$_2$CH$_2$	13 000.0	–1.11
74	id	PhCH=CHCH$_2$	54 000.0	–1.73
75	id	PhCO	52 000.0	–1.72
76	id	H	26 000.0	–1.41
77	id		38 000.0	–1.58
78	id	C$_6$H$_{11}$CH$_2$	410.0	0.39
79	id	adamantylCH$_2$	24 000.0	–1.38
80		*p*-CH$_3$O Bzl	440.0	0.36
81	id	*p*-Cl Bzl	240.0	0.62
82	id	CH$_3$	6 800.0	–0.83

Table IV. Human in vitro inhibition of AChE by *N*-benzylpiperidine-benzisoxazoles and their log ($1/IC_{50}$) calculated with the CoMFA model based on both steric and electrostatic contributions.



Compound	R	X	Y	IC_{50} (nM)	log $(1/IC_{50})_{obs}$	log $(1/IC_{50})_{pred}$
1	H	O	$(CH_2)_2$	55.00	1.26	0.30
2	5-Me	O	$(CH_2)_2$	7.80	2.11	0.62
3	5,6-diMe	O	$(CH_2)_2$	5.80	2.24	0.19
4	5-OMe	O	$(CH_2)_2$	7.20	2.14	0.50
5	6-OMe	O	$(CH_2)_2$	8.30	2.08	0.23
6	7-OMe	O	$(CH_2)_2$	7.10	2.15	0.17
7	6-NHCOMe	O	$(CH_2)_2$	2.80	2.55	0.86
8	6-NHCOPh	O	$(CH_2)_2$	9.40	2.03	0.81
9	6-$NHSO_2Ph$	O	$(CH_2)_2$	14.00	1.85	0.30
10		O	$(CH_2)_2$	0.80	3.10	0.93
11	6-NH_2	O	$(CH_2)_2$	20.00	1.70	0.07
12	6-OH	O	$(CH_2)_2$	26.00	1.59	0.14
13	6-Br	O	$(CH_2)_2$	50.00	1.30	0.20
14	6-CN	O	$(CH_2)_2$	101.00	1.00	–0.01
15	6-$CONH_2$	O	$(CH_2)_2$	8.80	2.06	0.70
16	H	O	$(CH_2)_3$	900.00	0.05	–0.94
17	H	O	E-CH=CH	210.00	0.68	–0.01
18	H	O	O-CH_2	2 600.00	–0.41	–0.58
19	H	O	NH-CH_2	320.00	0.49	–0.07
20	H	O	NH-$(CH_2)_2$	810.00	0.09	–0.32
21	H	S	$(CH_2)_2$	99.00	1.05	0.22
22	H	CH=CH	$(CH_2)_2$	220.00	0.66	–0.24
23	H	N=CH	$(CH_2)_2$	340.00	0.47	–0.28
24	H	NH	$(CH_2)_2$	120.00	0.92	0.28
25		–	–	0.33	3.48	1.08
26		–	–	3.60	2.44	0.48
27		–	–	0.57	3.24	0.75
28		–	–	0.95	3.02	0.59
29		–	–	0.48	3.32	1.24

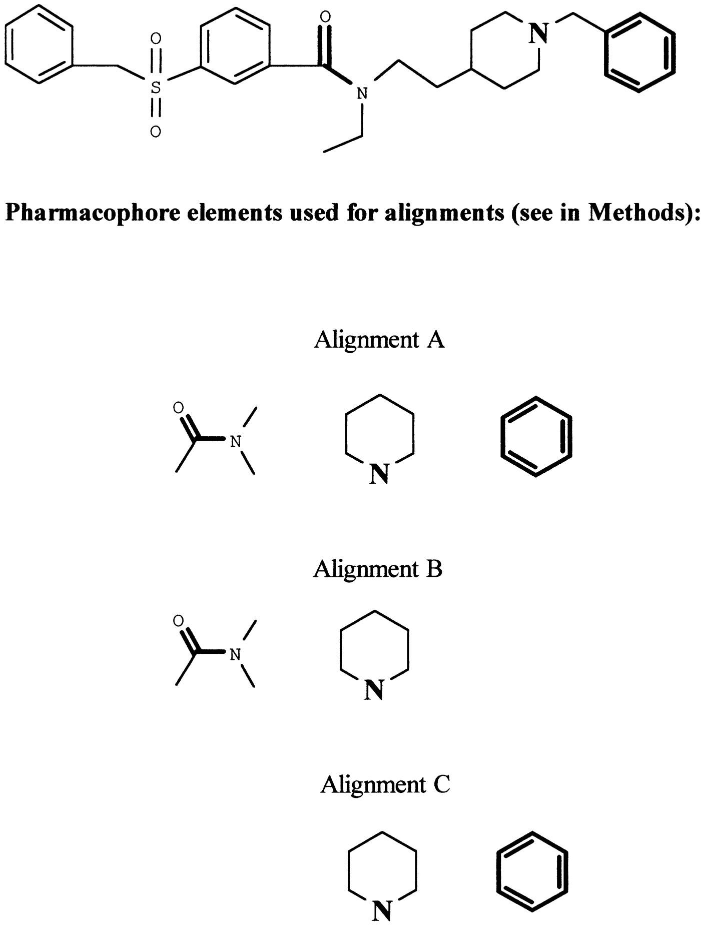

Pharmacophore elements used for alignments (see in Methods):
Alignment A
Alignment B
Alignment C

Figure 1. Compound **24** was used as reference for structural alignment of other compounds. Three pharmacophore elements used for the superposition are shown in bold.

derivatives (*tables I–III*) was able to predict the relative IC_{50} values for all the compounds of the test set (*table IV*). The following models were constructed.

– Assay model 4

The 82 compounds of the training set (*tables I–III*) and 29 compounds of the test set (*table IV*) in their AChE docked conformations were structurally superimposed according to alignment A (*figure 1*). The quaternary piperidine nitrogens were not protonated.

– Assay model 5

Calculations similar to those described above were performed with the protonated quaternary nitrogens of the piperidine moiety for the training and test sets. Alignment A (*figure 1*) was used for superimposing the molecules.

– Assay model 6

The same procedure was performed with the protonated pyridine nitrogens of compounds **14–16**, **29** and **38** from the training set (*table I*).

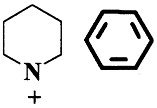
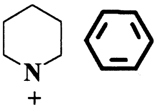
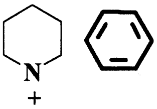
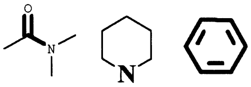
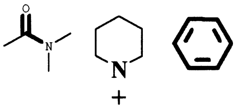
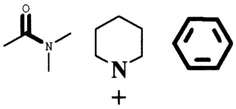
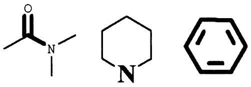
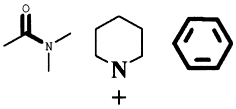
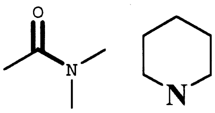
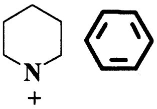
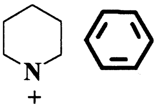
– Assay model 7

The conformations derived from the protein-based alignment were minimized thanks to the parameters mentioned above. The molecules were superimposed according to alignment A (*figure 1*). The piperidine nitrogens were not protonated.

– Assay model 8

The same calculations were performed with the protonated piperidine nitrogens.

Table V. Statistics of 3D QSAR models based on structural alignments (see details in methods). At the bottom of the table the corresponding results are given for protein-based alignment 3D QSAR models.

QSAR Model (see in Methods)	Number of compounds in the training set	Number of principal components	Q^2	R^2	F	SD	Pharmacophore elements used in superposition	Notes
1	66	5	0.57	0.91	117	0.40		every fifth compound excluded.
2	53	7	0.57	0.96	151	0.24		every fifth compound excluded.
3	53	8	0.59	0.97	170	0.22		every fifth compound excluded. Compounds **14**, **15**, **16**, **28**, **39** pyridines protonated.
4	82	5	0.61	0.90	136	0.42		
5	82	5	0.61	0.90	132	0.42		
6	82	6	0.63	0.92	140	0.38		compounds **14**, **15**, **16**, **28**, **39** pyridines protonated.
7	82	7	0.59	0.97	288	0.25		
8	82	8	0.55	0.97	274	0.24		
9	82	6	0.55	0.97	153	0.36		
10	66	5	0.58	0.91	117	0.40		
11	66	8	0.68	0.97	208	0.22		compounds **14**, **15**, **16**, **28**, **39** pyridines protonated.
[9]	82	7	0.75	0.98	508	0.19	protein-based alignment	docked to AChE active site

– Assay model 9.

The minimized conformations of the molecules of the training set were superimposed according to alignment B (*figure 1*). The compounds were not protonated.

– Assay model 10.

The 66 compounds from *tables I* and *II* served as the training set and the 29 compounds from *table IV* served as the test set, as before. These compounds were superimposed according to alignment C (*figure 1*). The quaternary nitrogens were protonated.

– Assay model 11.

The same procedure was performed with the protonated pyridine nitrogens of compounds **14–16**, **29** and **38** from the training set (*table I*).

2.6. Molecular diversity analysis

Kohonen SOM [22–24] was applied to investigate the molecular diversity of the training and test sets dealing respectively with the 82 (*tables I–III*) and with the 29 (*table IV*) N-benzylpiperidines. For this investigation, the following molecular descriptors were used: a series of 2D molecular descriptors, including 20 Kier-Hall molecular connectivity indices [25], ${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^3\chi_C$, ${}^3\chi_P$, ${}^4\chi_P$, ${}^4\chi_{PC}$, ${}^5\chi_P$, ${}^5\chi_C$, ${}^6\chi_P$, ${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$, ${}^3\chi^v{}_C$, ${}^3\chi^v{}_P$, ${}^4\chi^v{}_P$, ${}^4\chi^v{}_{PC}$, ${}^5\chi^v{}_P$, ${}^5\chi^v{}_C$, ${}^6\chi^v{}_P$, the number of paths and vertices with degrees 1–4, Gutman [26] and Platt indices, a series of information indices (IC^0, SIC^0, CIC^0, IC^1, SIC^1, CIC^1, IDW) [27, 28], the number of N, O and S atoms in a molecule, as well as a set of physico-chemical parameters, such as molecular weight, molecular volume, molecular refractivity, octanol-water partition coefficient [29] and a set of electronegativity parameters; i.e. that of a molecule by Sanderson [30] the mean, variance and maximum values of the atom electronegativity. The descriptors were normalized according to the following formula:

$$X^n_{ij} = \frac{Xij - Xj,\mathrm{min}}{Xj,\mathrm{max} - Xj,\mathrm{min}}$$

where *Xij* and $X^n{}_{ij}$ are the non-normalized and normalized j-th descriptor values for compound i, correspondingly, *Xj*, min and *Xj*, max are the minimum and maximum values for the j-th descriptor, and the j-th descriptor for compound i is the normalized descriptor value. The parameters for the Kohonen SOM procedure were set to the following values: size of the map: 10 × 10; number of learning iteration steps: 20 000; starting learning coefficient: 0.9. To increase the resolution of the map an interpolation option was used. Calculations were performed using the Neural-Works II software [23].

3. Results and discussion

The results of the calculations according to the alignments and the assay models described in the previous section are reported in *table V*. In the last row, the statistical criteria obtained in [14] for PBA 3D QSAR model are given. The number of principal components corresponds to the models with the smallest standard deviations in the cross-validation procedure [11], Q^2 is the cross-validated correlation coefficient, R^2 is the conventional correlation coefficient, F is the Fisher criterion and SD is the standard deviation. It is clearly seen that SBA QSAR models are satisfactory ($Q^2 > 0.5$), yet the results are not so good as with PBA QSAR models as to the prediction of the inhibition activity of the compounds belonging to the training set. As described in the previous section, models 1–3 were validated on a series of test compounds belonging to the same structural subgroups as in the training set. Statistics of these predictions are presented in *table VI*. The best prediction capability appeared in model 3. A graph of the experimental log ($1/IC_{50}$) values versus those predicted by model 3 for the compounds belonging to the test set is shown in *figure 2*. The only outlying compound, i.e. **30**, possesses a cyclohexyl group in the place where others have an aromatic ring or a short alkyl group. After removing this outlier, the following statistics were obtained: R = 0.97, SD = 0.24, F = 170. Therefore, the 3D QSAR models based on structural alignment can be used for the prediction of the inhibition activity of compounds belonging to the same subgroups of compounds that are presented in the training set. This is in agreement with results obtained by Welsh et al. [31], who used training and test sets including N-benzyl-piperidine derivatives with similar structures. Assay models 4–11 (see previous section) were applied to predict the activities of 29 N-benzyl-piperidine benzisoxazole or isoxazole tricycle derivatives (*table IV*), but R^2 values were inferior to 0.20 for all SBA models, whereas $R^2 = 0.90$ for the PBA model [14]. All these attempts can be considered as unsuccessful. So it is shown that contrary to the PBA 3D QSAR model [14], none of the

Table VI. Statistics of prediction of ID_{50} values for the compounds belonging to the test sets for 3D QSAR models 1–3. These compounds belong to the same subgroups presented in the training set.

Model	R^2	SD	F
1	0.69	0.81	31
2	0.64	0.72	19
3	0.73	0.64	28

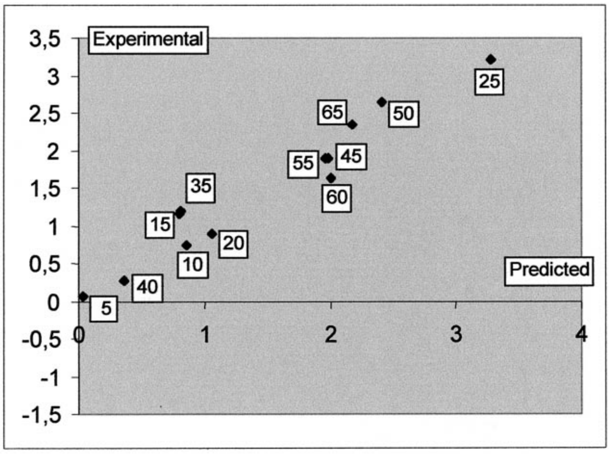


Figure 2. Experimental versus predicted log (1/IC$_{50}$) values for the test set compounds by the 3D QSAR model 3 values (the compound **30** was suppressed). The statistical criteria are R = 0.97, SD = 0.24, F = 170.

SBA 3D QSAR models was able to predict correctly the activities of the compounds from these groups.

3.1. Why 3D QSAR models based on protein alignment are better than structure-based models as regards to their predictive capabilities

When a compound is docked to a protein active site, side chains of the amino acid residues interact with the compound and both the compound and the side chains are spatially adjusted to each other. So each compound takes a slightly different position within the receptor's active site, and each of them has its own active conformation. These different conformations are superimposed by using a pharmacophore model, suggesting that some atoms and atom groups must have the same positions within the active site. This assumption is more or less inexact or even incorrect. This constraint has to be related to the reduced quality of structure-based 3D QSAR models as regards to their predictive power.

In *figure 3*, compound **24** (in red) in the training set and compound **10** (in violet, one of the most potent inhibitors in the test set), are shown in the AchE docked alignment in relation to each other. The same compound **10** (in brown), is shown in its position relative to compound **24** according to model 4. Compound **24** was the compound used as a reference for all the structure-based models considered in this study. The alignment for model 4 was constructed taking into account all three pharmacophore elements, so it was expected that in this case the structure-based alignment would be similar to the protein-based one. But this assumption, as it is clearly seen from *figure 3*, is incorrect. The distance between the amide nitrogens of compounds **24** in the training set, and **10** in the test set, for instance in the protein-based alignment is 2.2 Å, whereas in structure-based alignment 4 it is 0.84 Å and the corresponding distances between the amide oxygens are 4.67 Å and 2.25 Å, respectively. Moreover, in the structure-based alignment, the benzisoxazol moiety of compound **10** with a bulky substituent in position 6 occupies the area unfavorable for bulky substituents (*figure 3*), represented by a yellow surface. So, according to this model, the activity prediction for this compound must be unsatisfactory.

In models 5 and 6 (see Methods) the same conformations were used, so it is not surprising that the models constructed from these calculations were devoid of predictive power as regards to the external set of 29 benzisoxasols. The minimized conformations in models 6 and 7, as was shown using the best fit procedure, did not appear very different from those of the protein-based alignment, the rms values for most compounds with and without a proton at the quaternary nitrogen did not exceed 0.5 Å and 0.9 Å, respectively. So the reason why these models are not applicable for activity prediction is the same as in the two previous cases. The alignment for model 9 included only the quaternary nitrogen and amide (or the corresponding atoms), and the difference between the protein- and structure-based alignments appeared to be even larger than for models 4–8, so a good predictive power could hardly be expected from this model. As far as models 10 and 11 are concerned, the steric factor probably also plays a crucial role in their lack of predictive power for the 29 compounds of the test set.

A SOM neural network was used in this study in an attempt to elucidate the limitations of SBA models and classify the N-benzyl-piperidine derivatives involved in the investigation using a set of descriptors defined in the Method section. The results, presented in *figure 4,* will be discussed according to molecular diversity. All the compounds from the training and test sets were included in these calculations. In *figure 4a* the black diamonds and orange squares represent points corresponding to the compounds from the training set (*tables I–III*) and test set, respectively. It is clearly seen that both sets occupy different areas of the map, which means the compounds belonging to the training and test sets are different as regards their molecular diversity. Only two outlying compounds from the test set can be found in the other areas, i.e. compounds **20** and **28**. Compound **20** differs from the series of other compounds only by a substituent in the Y position (*table IV*). It is in the left lower corner and compound **28** is in the right lower corner of the map.

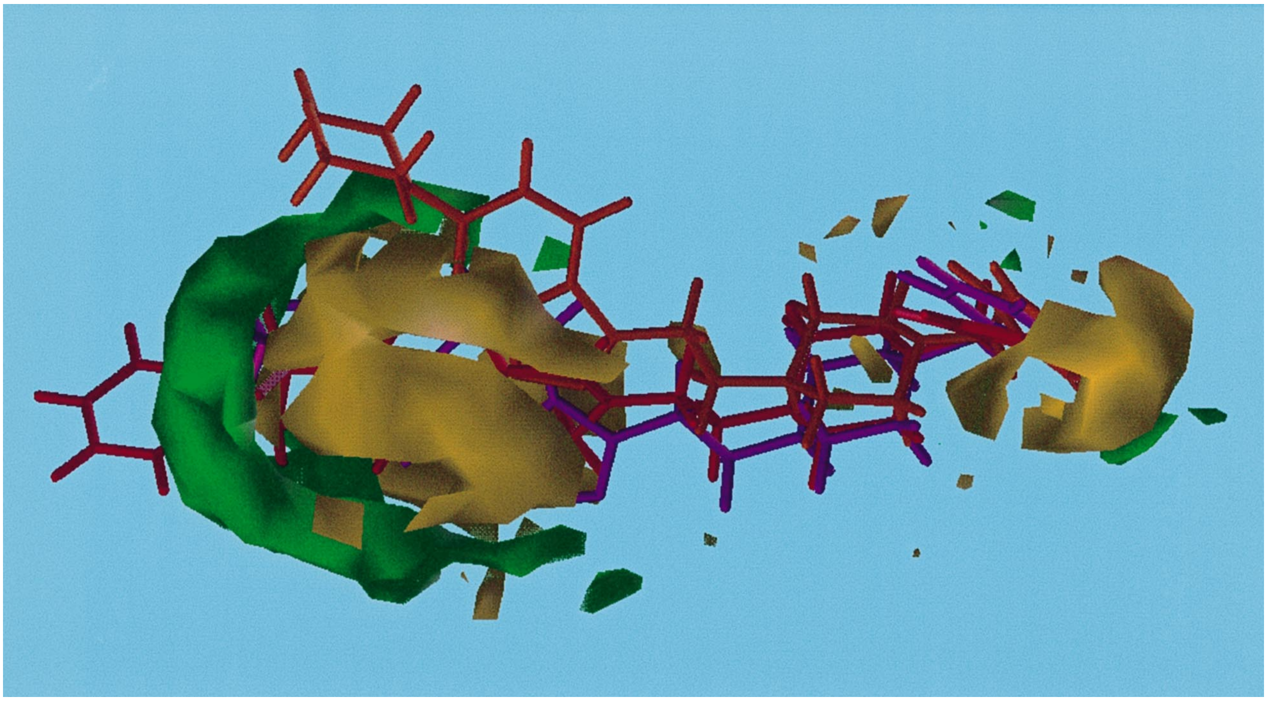


Figure 3. Compound **10** of the test set (violet and brown) is shown in accordance with protein-based and structure-based (model 4) alignments, correspondingly, together with compound **24** of the training set (red). Regions favourable and unfavourable for bulky substituents (according to model 4) are shown in green and yellow, respectively.

In the neighborhood of compound **28** the following compounds of the training set can be found: **23**, **24**, **25** as well as **13** (see also *table I*), all four compounds possessing *p*-($PhCH_2SO_2$)PhCO in position R_1 (*table I*). Compounds **23–25** are the most active in the training set, and compound **28** is one of the most active in the test set. This is not a coincidence, in *figure 4b* the same map is shown as in *figure 4a*, but the points in it are represented differently, according to the level of activity of the compounds. The blue diamonds correspond to the inactive compounds with log ($1/IC_{50}$) values inferior to 0.0 for the training set and to 0.5 for the test set. The green squares correspond to the compounds possessing an average activity with log ($1/IC_{50}$) values between 0.0 and 1.8 for the training set and between 0.5 and 2.0 for the test set. The yellow circles correspond to log ($1/IC_{50}$) values exceeding 1.8 for the training set and 2.0 for the test set. It is seen that active compounds occupy mainly the lower part on the right of the map, while the inactive ones spread throughout it. The latter results corroborate our previous findings that Kohonen SOM neural networks can be applied to investigate the molecular diversity of molecular databases and search for new leads [14, 24, 32, 33]. Here we emphasize, however, the other point that concerns the topic of this study, i.e. the impossibility to predict properties by the SBA 3D QSAR models of N-benzyl-piperidines which can be explained by molecular differences in the compounds belonging to the training and test sets. To demonstrate this, and to show the opposite is true for compounds similar to those from the training set, the distribution of the points corresponding to the 66 compounds belonging to the training set (*tables I* and *II*) is shown on the map in *figure 4c*. Compounds numbered **5**, **10**, **15**, up to **65** which belong to the test set for models 2 and 3 are represented by orange squares, the other compounds belong to the training set for these models and are represented by black diamonds. It is seen that in the vicinity of each orange square, a black diamond can be found. This means that there is little structural difference between the compounds belonging to both sets, with only one exception for compound **30**. At the same time, the results presented earlier in this paper as to the predictive capabilities of these models (see also *figure 2*) fully correspond to these findings.

At the same time, the strength of PBA 3D QSAR models is that they take into account the molecular diversity between the compounds in an indirect way. This

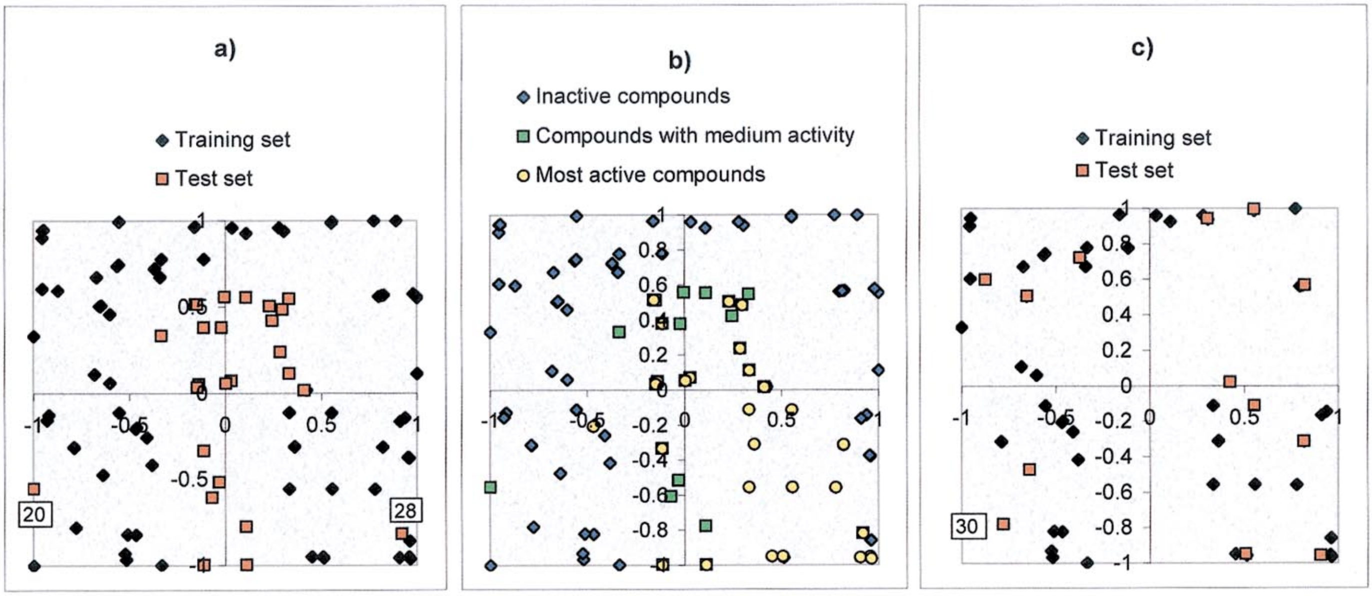


Figure 4. Study of molecular diversity between training and test sets using Kohonen SOM. **a**: points representing compounds belonging to the training and test sets for models 4–11. **b**: points corresponding to the same compounds shown in the map **a**, but represented according to the level of activity. **c**: points representing compounds belonging to the training sets for models 2 and 3.

is done with the help of the exact docked position of a compound, within the active site of the receptor, in accordance with the particular structure of this compound. So, the greatest advantage of the CoMFA model from [14] is its higher predictive power for compounds structurally different from those contained in the training set. Here it has been shown that it is hardly achievable by models based on structural alignment. The use of the structure-based models is strictly limited to compounds similar in terms of molecular diversity analysis to those in the training set. It demonstrates that, for good activity prediction, the PBA 3D QSAR models must be preferable to the SBA ones, whenever the 3D crystallographic study of the protein–ligand complex is available.

4. Conclusion

Two 3D QSAR approaches, i.e. a new one based on the PBA of ligands and one based on the SBA, were compared with regard to the predictive capabilities of the established models. The models were constructed for a series of N-benzyl-piperidines AChE inhibitors, most of them containing a benzoyl or a phtalimide moiety. The other models were constructed from eleven different structure-based 3D QSAR model assays with different pharmacophore elements for the superposition of molecules (*tables I–IV*) and states of piperidine and aromatic nitrogens, protonated or not protonated. The models were validated with log (1/IC$_{50}$) values, i.e. the inhibition activity of the compound, using the leave-one-out cross-validation procedure and tested on sets of compounds structurally more or less similar to those in the training set, or partially different from them, for example N-benzyl-piperidines.

It has been shown that 3D QSAR models based on structural alignment can be used to predict the inhibition activities of compounds belonging to the same subgroups of compounds that are in the training set. SBA 3D QSAR models contrast with the PBA ones insofar as none of them was able to correctly predict the activities of the compounds from the test set containing 29 N-benzyl-piperidines with structures partially dissimilar to those of the compounds contained in the training set.

By applying a molecular diversity analysis with the help of an SOM neural network, the compounds contained in the training and test sets, and structure-based models 4–11 were corroborated. So, the impossibility to predict properties by the structure-based 3D QSAR models of N-benzylpiperidines can be explained by molecular differences in the compounds belonging to the training and test sets due to the benzyisoxasole moiety of the last ones.

Thus the most interesting result consists of the high predictive power of the PBA CoMFA model developed to deal with compounds structurally different from those contained in the training set. This result is hardly achievable by models based on structural alignment, the use of which is limited to compounds similar to those of the training set. Then, CoMFA models based on ligand PBA must be preferred to structure-based models for biological activity predictions.

More generally, up to now, SBA is more widely used in CoMFA studies for the establishment of 3D QSAR models. But predictions must be restricted to the area of the molecular diversity of the training set as determined by self organizing maps (SOM). PBA is a more robust method, suitable outside the molecular diversity area of the training set, insofar that a suitable 3D crystallographic study of a protein–ligand complex is available.

Acknowledgements

Dr Marco Pintore is acknowledged for his valuable contribution in preparing the manuscript.

References

[1] SYBYL, Ligand-Based Design Manual, Tripos, Inc. (1996).

[2] Burkert U., Allinger N.L. (Eds.), Molecular Mechanics, ACS, Washington DC, 1982.

[3] Cramer III R.D., Patterson D.E., Bunce J.D., J. Amer. Chem. Soc. 110 (1988) 5959–5967.

[4] Marshall G.R., Cramer III R.D., Trends Pharmacol. Sci. 9 (1988) 285–289.

[5] Clementi S., Wold S., in: van de Waterbeemd H. (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 319–338.

[6] Wold H., in: Gani J. (Ed.), Perspectives in Probability and Statistics, Academic Press, London, 1975.

[7] Wold H., in: Johnson N.L., Kotz S. (Eds.), Encyclopedia of Statistical Sciences, J. Wiley and Sons, New York, 1984.

[8] Wold S., in: van de Waterbeemd H. (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 195–218.

[9] Adams M.J., Chemometrics in Analytical Spectroscopy, (Ed.), The Royal Society of Chemistry, Cambridge, 1995.

[10] Wold S., Eriksson L., in: van de Waterbeemd H. (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 309–318.

[11] Kubiny H., Quant. Struct.-Act. Relat. 13 (1994) 285–294.

[12] Brennan M.B., Chem. Eng. News 20 (1997) 29–35.

[13] John V., Lieberburg I., Thorsett E.D., Annu. Rep. Med. Chem. 28 (1993) 197–206.

[14] Bernard P., Kireev D.B., Chrétien J.R., Fortier P.L., Coppet L., J. Comput.-Aided Mol. Des. 13 (1999) 355–371.

[15] Rarey M., Kramer B., Lengauer T.J., J. Comput.-Aided Mol. Des. 11 (1997) 369–384.

[16] Makino S., Kuntz I.D., J. Comput. Chem. 18 (1997) 1812–1825.

[17] Welch W., Ruppert J., Jain A.N., Chem. Biol. 3 (1996) 449–462.

[18] Harel M., Schalk I., Ehret-Sabatier L., Bouet F., Goeldner M., Hirth C. et al., Proc. Natl. Acad. Sci. USA 90 (1993) 9031–9035.

[19] Austel V., in: van de Waterbeemd H. (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 49–62.

[20] Sugimoto H., Tsuchia T., Sugumi H., Higurashi K., Karibe N., Iimura Y. et al., J. Med. Chem. 33 (1990) 1880–1887.

[21] Sugimoto H., Tsuchia T., Sugumi H., Higurashi K., Karibe N., Iimura Y. et al., J. Med. Chem. 35 (1992) 4542–4548.

[22] Kohonen T. (Ed.), Self-organization and Associative Memory, Springer-Verlag, Berlin, 1988.

[23] Neural Computing. NeuralWare, Inc., (1995).

[24] Bernard P., Golbraikh A., Kireev D., Chrétien J.R., Rozhkova N., Analusis 26 (1998) 333–341.

[25] Kier L.B., Hall L.H. (Eds.), Molecular Connectivity in Structure–Activity Analysis, John Wiley and Sons, New York, 1986.

[26] Gutman I., Ruscic B., Trinajstic N., Wilcox Jr C.F., J. Chem. Phys. 62 (1975) 3339–3405.

[27] Sabljic A., in: Karcher W., Devillers J. (Eds.), Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology, Kluwer Academic Publishers, Dordrecht, 1990, pp. 61–82.

[28] Basac S.C., in: Karcher W., Devillers J. (Eds.), Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology, Kluwer Academic Publishers, Dordrecht, 1990, pp. 83–103.

[29] Hansch C., Leo A. (Eds.), Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley Interscience Publication, New York, 1979, p. 13.

[30] Sanderson R.T. (Ed.), Chemical Bonds and Bond Energy, Academic Press, New York, 1976.

[31] Tong W., Collantes E.R., Chen Y.U., Welsh W.J., J. Med. Chem. 39 (1996) 380–387.

[32] Kireev D.B., Ros F., Bernard P., Chrétien J.R., Rozhkova J.R., in: van de Waterbeembd H., Testa B., Folkers G. (Eds.), Computer-Assisted Lead Finding and Optimization, Verlag Helvetica Chimica Acta, Basel and Wiley-VCH, Weinheim, 1997, pp. 255–264.

[33] Kireev D.B., Chrétien J.R., Bernard P., Ros F., SAR and QSAR in Environmental Research 8 (1998) 93–107.