Available online at www.sciencedirect.com

SCIENCE DIRECT®

Bioorganic & Medicinal Chemistry

Bioorganic & Medicinal Chemistry 14 (2006) 1348–1363

Affinity prediction on A_1 adenosine receptor agonists: The chemometric approach

Paola Fossa,[a,*] Luisa Mosti,[a] Francesco Bondavalli,[a] Silvia Schenone,[a] Angelo Ranise,[a] Chiara Casolino[b] and Michele Forina[b]

[a]*Dipartimento di Scienze Farmaceutiche, Università degli Studi di Genova, Viale Benedetto XV 3, I-16132 Genoa, Italy*

[b]*Dipartimento di Chimica e Tecnologie Farmaceutiche e Alimentari, Università degli Studi di Genova,*
Via Brigata Salerno, I-16147 Genoa, Italy

Received 17 February 2005; revised 8 September 2005; accepted 27 September 2005
Available online 2 November 2005

Abstract—In this paper, we are presenting a quantitative-structure–activity relationship (QSAR) study performed on 21 selective A_1 adenosine receptor agonists plus the endogenous substrate, adenosine, so as to identify those predictors which play a key role in describing the binding of the ligand with the A_1 receptor. A large number of molecular descriptors plus a calculated receptor–agonist binding energy and atomic charges were taken into account to derive different QSAR models, using different regression techniques. The results obtained both with linear and nonlinear approaches converge to the selection of the same informative parameters, highlighting the correlation of these descriptors with the biological Response. The evaluation 'a priori' of these predictors could therefore represent a useful tool in the screening of large libraries of compounds and in the rational design of new selective agonists.

1. Introduction

Adenosine is a ubiquitous neuromodulator in both the peripheral and the central nervous systems. The effects elicited by adenosine are mediated by its interactions with four receptor subtypes termed A_1, A_{2A}, A_{2B}, and A_3, which can be distinguished pharmacologically, based on the rank order of the potency of agonists and antagonists. These receptors belong to the superfamily of G protein-coupled receptors and contain seven transmembrane domains (α-helices), interconnecting loops, an extracellular terminal amino residue, and a cytoplasmic terminal carboxylate residue.

Adenosine receptors (AR) from different species show a very high amino acid sequence homology (82–93%), with the only exception of the A_3 subtype which exhibits a 74% primary sequence homology between rat and human or sheep.[1]

Many selective agents have been developed until now for the A_1 receptor subtypes and some of these seem promising as potential therapeutic agents in the treatment of neurodegenerative disorders;[2] however, there are currently very few A_1 adenosine ligands under clinical trial.

In this context, as a further development of our research in understanding the structural basis of ligand–A_1 receptor interactions,[3,4] we have focused our interest in identifying the molecular properties responsible for affinity toward A_1 AR by means of a quantitative-structure–activity relationship (QSAR) study on a set of A_1 agonists. As is well known, in these methods correlations are derived between experimentally determined binding affinities and a number of different descriptors, which should variously encode for the thermodynamics of binding of a set of ligands. The base assumption, in fact, is that a correlation exists between the enthalpy of binding of the molecule and its molecular properties. A set of 21 selective A_1 agonists **2**–**22** (Fig. 1) selected from the literature, plus adenosine, **1**, the natural endogenous agonist of the receptor, was thus considered.[5]

For the prediction, the structural information encoded by a large number of molecular descriptors for topological, electronic, geometric, and polar surface properties, derived from DRAGON,[6] was taken into account together with atomic charges on those specific positions of the adenosine skeleton, highlighted as important by

Keywords: QSAR; A_1 adenosine agonists; Chemometric approach.

* Corresponding author. Tel.: +39 10 3538361; fax: +39 10 3538358; e-mail: fossap@unige.it

0968-0896/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2005.09.058

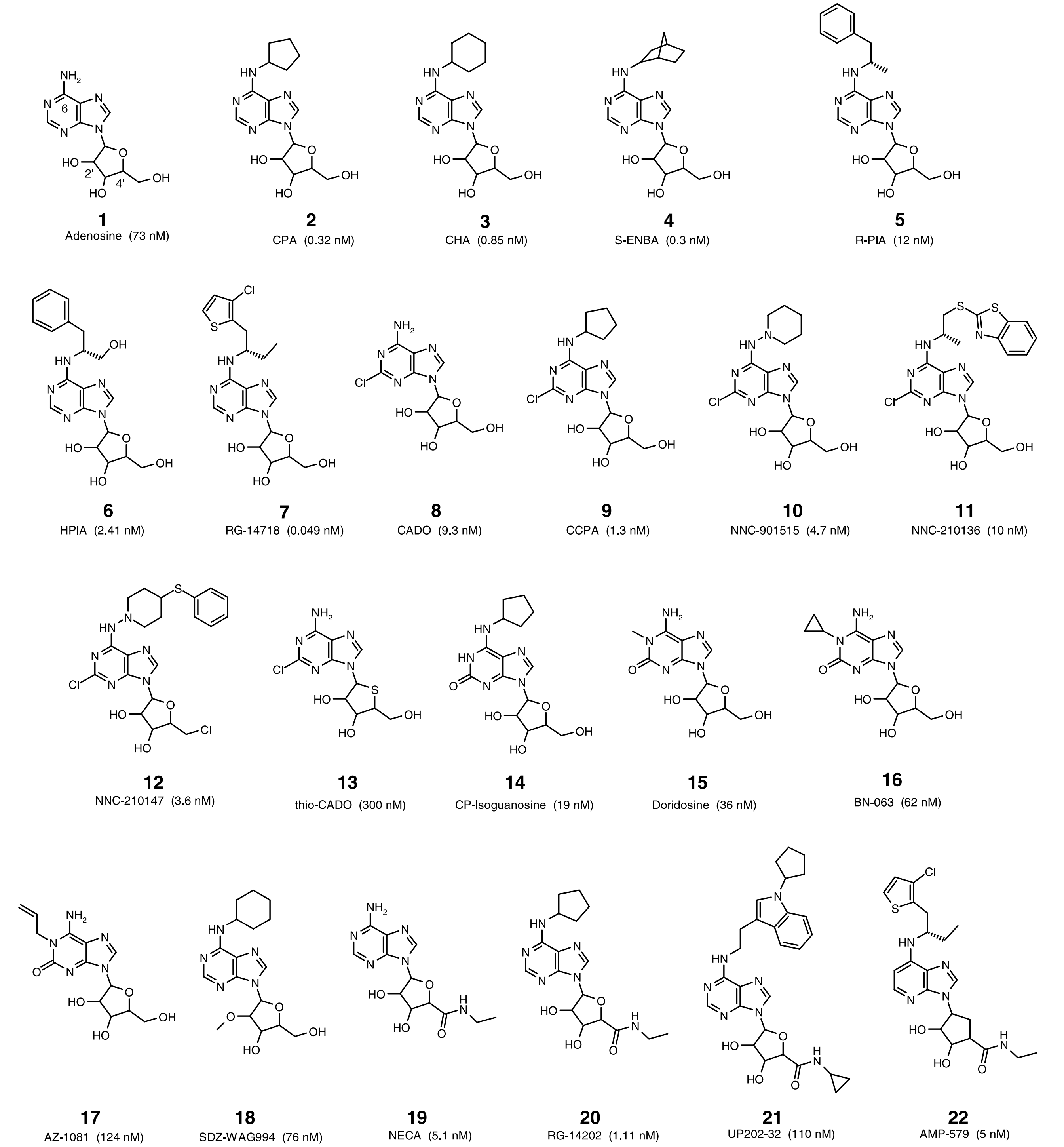

Figure 1. Molecular formulas of adenosine **1** and selective A_1 agonists **2**–**22**. For each structure, K_i values determined on A_1 AR from rat brain membranes and expressed as nM are given in parentheses [Ref. 5 and literature cited therein].

previous structure–activity relationship studies.[5a] In addition, calculated receptor–ligand binding energies for compounds **1**–**22** were included among variables. The versatile chemometric package PARVUS[7,8] was subsequently applied to handle such information, discarding all non-informative descriptors and extracting meaningful QSAR models able to identify the most important predictors accounting for affinity toward A_1 receptor.

2. Methods

2.1. Data set

For this investigation, 21 selective A_1 AR agonists **2**–**22** plus adenosine **1** were collected from the literature. Their molecular structure and their affinity value toward the A_1 adenosine receptor from rat cerebral cortex, expressed as K_i (nM), are reported in Figure 1. The reduced availability of affinity data on A_1 agonists for the human A_1 receptor hampers the study of agonist–receptor interactions in the A_1AR model, that is why we employed a larger data set of affinity values as determined in rat brain, considering that the sequences of the A_1AR in the two species have a percentage identity of 94.8% and the amino acid differences are located in the TM4-5 loop and in the carboxy-terminal cytoplasmic segment.

The synthesis and the biological data of compounds **1**–**22** are reported in Ref. 5. Calculations were performed transforming the original affinity into:

$$\text{Affinity} = -\log(1 + K_i)$$

and using it as Response variable.

With this transform the distribution of the activities is almost uniform, without high leverage points, as observed in the case of K_i histogram, and in any case better than that obtained using the transform $-\log K_i$ (pK_i), very commonly employed in QSAR studies (Fig. 2).

The docked energy conformations of compounds **1**–**22** previously determined by us[4] were used as input files for DRAGON and Spartan. All calculations were run on personal computers.

2.2. Descriptors generation

For each compound the package DRAGON, version 3.0, was used to calculate 1481 molecular descriptors grouped as follows: constitutional descriptors (47), topological descriptors (262), molecular walk counts (21), BCUT descriptors (64), Galvez topological charge indices (21), 2D autocorrelations (96), charge descriptors (14), aromaticity indices (4), Randic molecular profiles (41), geometrical descriptors (58), RDF descriptors (150), 3D-MoRSE descriptors (160), WHIM descriptors (99), GETAWAY descriptors (197), functional group descriptors (121), atom-centered descriptors (120), empirical descriptors (3), and properties (3). By means of the 'variable exclusion' option in DRAGON, constant variables, near constant variables, and correlated pairs were discarded, leaving 730 descriptors. Furthermore, since numerous studies have highlighted the importance of substitution on specific positions of adenosine skeleton (Ref. 5a and literature cited therein), atomic charges, determined with semi-empirical AM1 calculations using Spartan 'O2[9], on positions C2, C6, N6, X2 (atom of the substituent on position 2 connected to C2), 2′-ribose, 4′-ribose, and *O*-ribose atom, were included. Calculated receptor–ligand binding energies for **1**–**22**, determined according to the procedure previously described,[4] were also added. A final matrix of 22 objects and 739 rows was thus obtained.

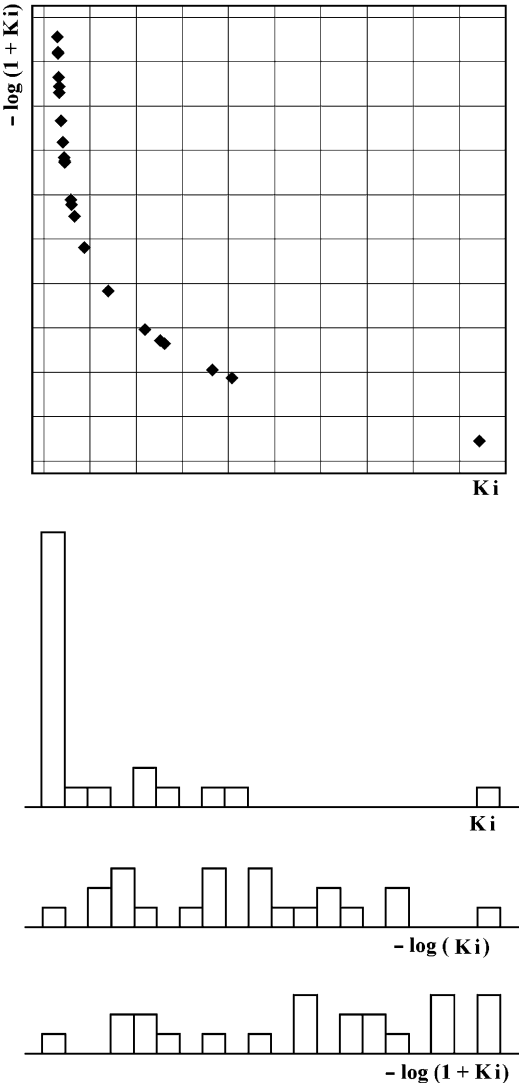

Figure 2. Diagram of K_i versus $-\log(1 + K_i)$ and histograms of the distribution of K_i, $-\log K_i$, and $-\log(1 + K_i)$ values for compounds **1**–**22**.

Table 1. Cut-off correlation coefficient and predictors retained in the 11 X-Blocks

Index	Block	r	Retained predictors
1	A	0.99	332
2	B	0.95	173
3	C	0.90	112
4	D	0.85	76
5	E	0.825	60
6	F	0.81	55
7	G	0.80	50
8	H	0.775	43
9	I	0.75	40
10	L	0.725	33
11	M	0.70	28

2.3. Variables selection

A more accurate selection of the information encoded by the selected variables was carried out by means of the CHECK module implemented in PARVUS. Eleven blocks of predictors have been extracted from the original data. The first block (X-block A, Table 1) contains 332 predictors. It was obtained by a preliminary selection technique based on the elimination of: (i) constant predictors, (ii) predictor constants in one of the training sets of cross-validation (CV) performed with five cancellation groups, (iii) one of two predictors with linear correlation coefficient $r \geqslant 0.99$, (iv) the predictors that have less than four levels so that they cannot be used in cubic regression.

Other 10 X-blocks (B-M) were obtained with the same procedure, with a different cut-off value for the correlation coefficient, from 0.95 to 0.7, as shown in Table 1.

As supplementary material available from authors, Appendix I shows the elimination of the predictors in detail. For example, predictor 3 (N6) was eliminated in block H because its correlation coefficient with predictor 2 (C6) was larger than the cut-off value for example of 0.775. Appendix II lists and briefly describes all predictors used in this study.

2.4. Chemometric strategies

To evaluate the relevance of the selected predictors with regard to the possibility of prediction of the biologic activity, the following techniques have been applied:

Partial least-squares (PLS) regression.[10,11]
Stepwise ordinary least-squares (SOLS) regression.[10]
Iterative predictor weighting PLS (IPWPLS).[12]
Iterative stepwise elimination PLS (ISEPLS).[13]
Uninformative variable elimination PLS (UVEPLS).[14]
Generating optimal linear PLS Estimations (GOLPE).[15,16]
Quadratic stepwise regression (QSR).
Alternating conditional expectations (ACE) regression.[17]

Moreover, in the development of data analysis, we also used principal component analysis (PCA)[10,11] and clustering analysis[18] with seriation.[19]

All these techniques are used as implemented in PARVUS,[7,8] where they are included with minor differences in comparison to the original algorithms.

Only the first regression technique works with all the predictors; the other techniques search for a subset of relevant predictors under the hypothesis of linear relationships (SOLS, IPWPLS, ISEPLS, UVEPLS, and GOLPE) or of nonlinear relationships (QSR, ACE).

In the above techniques, the strategies used to select the relevant predictors are very different. The performances of these techniques have never been compared before, only some papers employing two or three techniques are presented in the literature. For the selection of useful predictors many other techniques are however available, as those based on genetic algorithms (GA)[20] used with OLS or PLS[21] or those based on modified techniques as LASSO.[22]

Thus, in this work many different selection techniques were used to search for their consensus regarding the most important predictors and in order to avoid the possibility of overestimating the prediction ability. GA selection was used only in the exploratory analysis. The results obtained varied greatly depending on the settable parameters (number of GA runs, number of epochs, elitism, mutation probability, maximum number of selectable predictors, and number of CV groups). The selected predictors became very different depending on the parameters; the complexity of the models was always being very large. In fact, GA selects the predictors optimizing the prediction ability which is always overestimated when the number of selected predictors and the complexity of PLS model approach the number of objects. Data were always autoscaled, that is, for each variable (predictors or Response) its mean was subtracted and the result was divided by the standard deviation. Mean and standard deviation were computed only with data in the training sets [11] and used to scale data in the evaluation sets. Validation was performed by means of CV[10] with the leave-one-out procedure or with five cancellation groups.

Cross-validation is generally a good method to evaluate model quality because it consists in the prediction of the response for each object using a model built without considering the information related to this object.

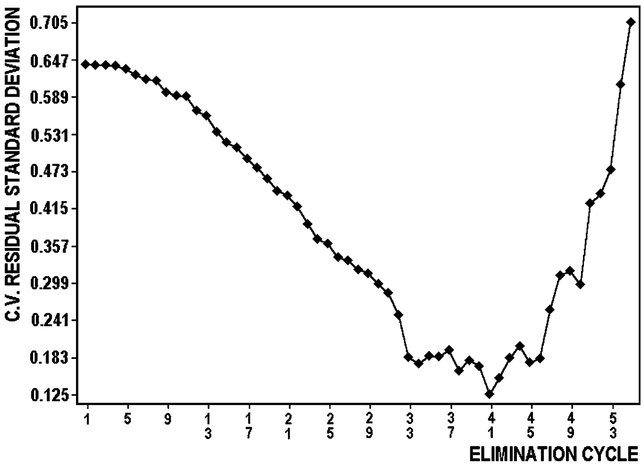

Figure 3. Example of the variation of the predictive ability in ISEPLS cycles (X-block A).

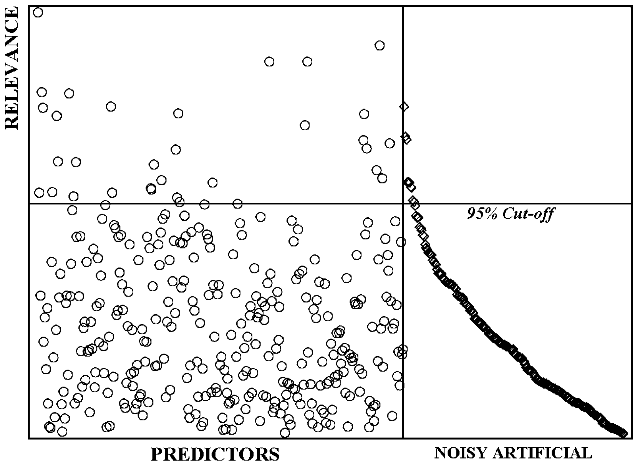

Figure 4. UVEPLS evaluation of the 332 predictors (X-block A) plus 200 noisy variables. Noisy added variables are ordered according their relevance.

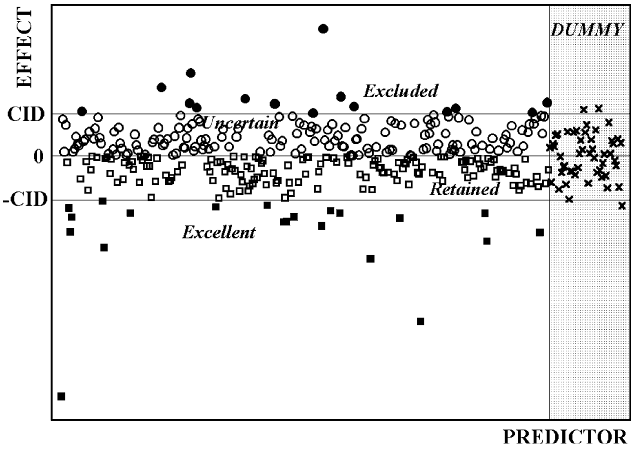

Figure 5. V-GOLPE evaluation of the 332 predictors (X-block A).

However, in some cases, calibration methods need optimization steps based on CV, and the prediction ability of the computed models becomes overestimated. Consequently, our study was enlarged creating a little external test set (four molecules). Due to the reduced number of available molecules compared to the predictor number, the external set was chosen in order to permit anyway an optimal construction of the calibration models. The training set molecules have been selected to spread throughout the entire predictor space. The space-filling Kennard and Stone[23] algorithm was applied using the first 10 principal components of the autoscaled data, considering 94% of total variance. Seventeen objects were selected for training set and the molecules 3 (CHA), 5 (R-PIA), 6 (HPIA), and 10 (NNC-901515) formed the test set. Calibration models have been studied on the remaining 18 molecules, optimized by means of cross-validation, and finally tested on the other four molecules totally stranger to the sequential steps of model building. All the different selection strategies were tested on this data set.

Table 2. Percent explained variance and cumulative explained variance with PCA applied to X-Block A plus the Response

Component	% Expl. variance	% Cumulative
1	66.00	66.00
2	7.85	73.85
3	5.46	79.31
4	3.29	82.60
5	2.97	85.57
6	2.47	88.04
7	2.11	90.14
8	1.74	91.88
9	1.16	93.05
10	1.09	94.13

In the case of our data matrices, it is impossible to apply ordinary least-squares regression (OLS), because the number of variables is larger than the number of objects, so the information matrix of OLS cannot be inverted. Thus, biased regression techniques have been applied.

PLS[10,11] regression has been widely applied to QSAR problems in the past few years. It can work with the complete matrix of the predictors. PLS computes 'latent variables,' linear combinations of the predictors. The first latent variable is the linear combination of predictors with the largest covariance with the Response variable. From the loadings of the predictors on the latent variable it is possible to obtain the so-called closed form of PLS, with provisional values of the b coefficients:

$$\hat{y} = b_0 + b_1x_1 + \cdots + b_\mathrm{v}x_\mathrm{v} + \cdots + b_\mathrm{V}x_\mathrm{V} = \mathbf{x}^\mathrm{T}\mathbf{b}. \quad (1)$$

To compute the second latent variable, only the information orthogonal to the first one is retained in the X-block of the predictors, and the vector of the

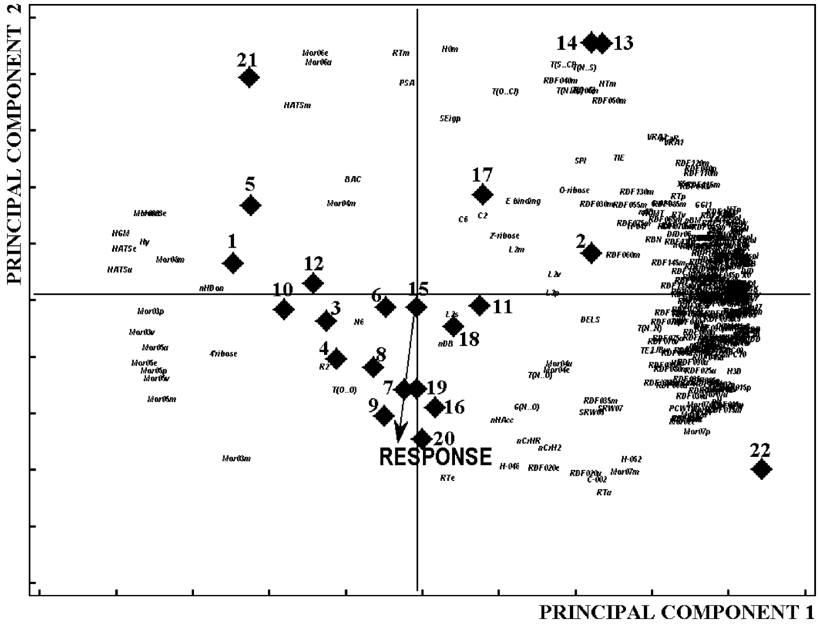

Figure 6. PC Biplot for X-block A joined with the Response. Autoscaled data.

Responses is substituted by the residuals, that is, the difference between the original Responses and the values estimated by Eq. 1. After the selection of the model complexity (number of latent variables in the final PLS model) corresponding to the minimum residual standard deviation of the prediction of the Response (SEP), the final values of the b coefficients of Eq. 1 are computed.

SEP, standard error of prediction, frequently indicated with other acronyms as SDEP for standard deviation of the error of prediction, or RMSECV for root-mean-square error of cross-validation, is measured on the evaluation sets of cross-validation:

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}, \tag{2}$$

where N is the number of objects.

SOLS[10,8] is probably the oldest biased regression technique. It first selects the predictor with the largest correlation coefficient with the Response, then the predictor with the largest correlation coefficient with the residuals, and so on. In each step, an F test is applied to evaluate the significance associated with the introduction of a new predictor in the model. The number of predictors selected is always small. So, the information matrix X^TX can be inverted and the usual algorithm of multiple linear regression can be applied.

IPWPLS,[12] ISEPLS,[13] UVEPLS,[14] and GOLPE[15,16] are techniques developed to eliminate the useless predictors, associated to the PLS algorithm. The first two methods found their elimination strategy on an iterative procedure. PLS regression is repeated many times gradually reducing the number of predictors depending on their importance, product between the absolute value of the regression coefficient b_v and the standard deviation s_v (1 in the case of autoscaled data). In each step, IPW modify the effect of each predictor weighting it by its importance, ISE simply deletes the less important ones (one or more predictors at time). Figure 3 shows how the predictive ability varies in ISE, applied on data set A by eliminating, in each cycle, the worst predictors, having less than 3% of the residual predictors eliminated in each cycle. The final model is the one with the maximum predictive ability.

UVEPLS[14] adds random variables, with very small value, to the original predictors and computes the reliability c_v of each predictor v, original or artificial, as the ratio between the corresponding PLS regression coefficient b_v and its standard deviation, s_{b_v}. The cut-off value for the elimination of non-informative original predictors is the maximum absolute value of the coefficient c_v for the added artificial predictor, or the value corresponding to their α% quantile (usually 95%). Figure 4 shows the results obtained by 95%-UVE on data set A.

GOLPE[15,16] found its elimination strategy on the study of a large number, M, of 'PLS reduced models,' obtained deleting some variables at time. The presence or absence of the predictors in each model is defined in a 'design matrix' built by means of fractional factorial design (FFD) scheme.

In this study, this method was used as implemented in V-PARVUS, as V-GOLPE. The design matrix is obtained by random generation (at a pre-selected

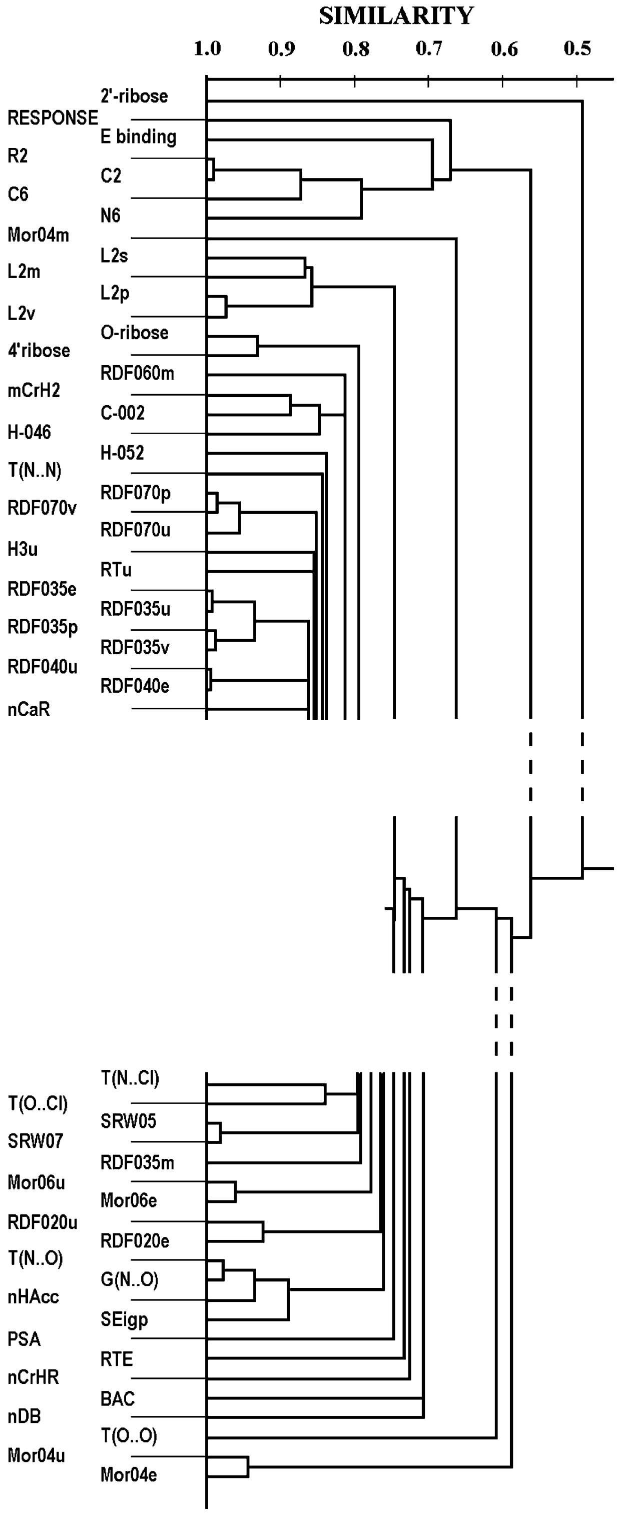

Figure 7. Dendrogram of similarities calculated for all the selected predictors. For clarity, only the extreme parts are reported. In the upper part are reported the descriptors more related to the independent variable Response, in the lower part the descriptors less related to it are listed.

Table 3. Results of regression techniques, grouped for X-block

X-Block	Technique	Predictors	Complexity	% Fitting R^2	SEC	% Prediction R^2 CV	SEP
A	PLS	332	4	94.74	0.169	43.71	0.576
	SOLS	6	6	95.30	0.189	90.21	0.237
	IPW	7	5	93.70	0.185	91.70	0.221
	ISE	24	6	99.67	0.043	95.62	0.158
	UVE	6	4	48.66	0.529	31.12	0.627
	UVE 99%	22	4	88.61	0.249	69.62	0.417
	UVE 95%	36	4	91.11	0.220	71.86	0.401
	UVE 90%	36	4	91.11	0.220	71.86	0.401
	QSR	3	6	89.80	0.236	86.72	0.280
	ACE	8		99.46	0.067	95.60	0.151
B	PLS	173	3	85.80	0.278	38.89	0.600
	SOLS	9	9	97.43	0.118	95.88	0.153
	IPW	7	5	93.69	0.186	91.72	0.221
	ISE	30	6	99.81	0.021	96.42	0.143
	UVE 95%	20	9	96.22	0.144	68.88	0.422
	QSR	5	10	91.85	0.211	91.38	0.225
	ACE	8		99.97	0.014	96.29	0.139
C	PLS	112	10	99.92	0.019	39.60	0.597
	SOLS	10	10	96.62	0.136	92.96	0.169
	IPW	13	9	98.65	0.086	96.94	0.134
	ISE	21	8	99.79	0.034	97.77	0.113
	UVE 95%	16	11	96.15	0.145	49.37	0.538
	QSR	5	10	91.85	0.211	91.38	0.225
	ACE	10		99.96	0.020	95.55	0.152
D	PLS	76	10	99.90	0.024	46.55	0.561
	SOLS	4	4	77.12	0.393	68.55	
	IPW	12	10	98.49	0.091	91.85	0.219
	ISE	19	10	99.81	0.033	98.53	0.092
	UVE 95%	11	7	91.80	0.212	76.06	0.370
	QSR	5	10	91.85	0.211	91.38	0.225
	ACE	10		99.97	0.017	96.05	0.143
E	PLS	60	13	99.98	0.011	53.18	0.525
	SOLS	4	4	77.12	0.393	68.55	0.340
	IPW	7	7	96.06	0.147	90.71	0.234
	ISE	30	6	98.86	0.079	85.30	0.290
	UVE 95%	7	2	64.28	0.441	49.02	0.540
	QSR	5	10	91.87	0.211	92.76	0.207
	ACE	10		99.94	0.025	96.16	0.141
F	PLS	55	9	99.84	0.030	47.58	0.556
	SOLS	4	4	77.11	0.393	68.55	0.424
	IPW	13	9	98.57	0.088	91.41	0.225
	ISE	17	8	99.60	0.047	97.16	0.127
	UVE 95%	10	2	81.55	0.317	70.60	0.410
	QSR	4	8	91.85	0.211	91.38	0.225
	ACE	9		99.79	0.043	91.11	0.215
G	PLS	50	9	99.84	0.029	49.92	0.543
	SOLS	4	4	76.76	0.396	64.64	0.333
	IPW	13	9	98.60	0.087	91.42	0.225
	ISE	17	8	99.60	0.047	97.16	0.127
	UVE 95%	8	2	70.24	0.403	44.27	0.564
	QSR	5	10	91.85	0.211	91.38	0.225
	ACE	8		99.72	0.048	88.63	0.243
H	PLS	44	10	99.81	0.032	50.00	0.543
	SOLS	4	4	76.76	0.396	64.64	0.333
	IPW	9	4	94.20	0.178	90.46	0.237
	ISE	17	9	99.06	0.072	96.21	0.147
	UVE 95%	5	1	62.31	0.453	49.56	0.537
	QSR	5	10	91.85	0.211	91.38	0.225
	ACE	8		99.51	0.063	86.94	0.261

Table 3 (*continued*)

X-Block	Technique	Predictors	Complexity	% Fitting R^2	SEC	% Prediction R^2 CV	SEP
I	PLS	40	10	99.75	0.037	49.83	0.540
	SOLS	4	4	76.76	0.252	64.64	0.333
	IPW	7	4	93.12	0.194	87.49	0.271
	ISE	17	10	99.65	0.044	97.63	0.116
	UVE 95%	6	2	59.56	0.470	47.95	0.545
	QSR	5	10	91.85	0.211	91.38	0.225
	ACE	8		99.69	0.050	85.82	0.272
L	PLS	33	9	99.49	0.053	51.71	0.533
	SOLS	4	4	76.76	0.396	64.64	0.333
	IPW	7	4	93.03	0.195	86.94	0.277
	ISE	18	9	98.94	0.076	96.49	0.142
	UVE 95%	3	2	48.66	0.529	40.37	0.584
	QSR	5	10	91.85	0.211	91.38	0.225
	ACE	7		99.69	0.048	88.95	0.240
M	PLS	28	2	76.31	0.359	33.78	0.625
	SOLS	4	4	73.06	0.426	20.62	0.674
	IPW	5	3	72.83	0.385	66.80	0.442
	ISE	14	1	71.89	0.392	54.21	0.512
	UVE 95%	4	1	51.63	0.514	46.20	0.555
	QSR	5	10	91.85	0.211	91.38	0.225
	ACE	10		99.83	0.040	91.41	0.211

probability level) of the predictor condition used-not used. M is as large as desired (not a power of 2 as in FFD) and it is possible to explore more combinations of used predictors.

The effect of each predictor v, E_v, is computed relating the prediction ability of the models (SDEP values obtained by means of cross-validation procedure) with the presence or the absence of the predictor. The obtained values are then compared to the effect of some dummy variables, introduced in the design matrix, by means of the CID, confidence interval of the effect of the dummies (95% value of the Student distribution), computed in V-GOLPE as:

$$\mathrm{CID} = \sqrt{\frac{\sum_{d=1}^{D}(E_d - \bar{E})^2}{D}}t_{\mathrm{crit}}, \tag{3}$$

where E_d is the effect of each dummy variable $\bar{E}$ is their mean effect (about 0 when D, the number of dummy variables is large). The number of dummy variables has no effect on the computer time.

From this comparison the predictors are labelled as excellent, retained, excluded or uncertain.

V-GOLPE computes PLS models after the elimination of Excluded predictors (Elimination level I), then after the elimination of Uncertain predictors (Elimination level II), and finally only with the Excellent predictors (Elimination level III). Figure 5 shows the results obtained by V-GOLPE on data set A.

QSR is a stepwise procedure that selects in each step the variable that best predicts (CV) Response values (or residual values) by means of least-squares quadratic regression, using a model as:

$$\hat{y} = b_0 + b_{1\mathrm{s}}x_\mathrm{s} + b_{2\mathrm{s}}x_\mathrm{s}^2. \tag{4}$$

The difference with SOLS is not only in the form (quadratic instead of linear) but also in the predictive evaluation of the best predictor in each step.

ACE[17] is a nonlinear regression method. It can be applied when the ratio between the number of objects and number of variables is very high (at least 10).

In the ACE model, the Response variable is the sum of smoothed functions of the predictors:

$$\hat{y} = b_0 + \sum_{v=1}^{V} t_v(x_v), \tag{5}$$

The smoothed functions $t_v\ (x_v)$ are not explicit functions of the predictors, as the usual transforms, squares, logarithms, roots used, for example, in nonlinear OLS. They are, instead, the result of ACE algorithm, and they are obtained as piecewise linear functions of the predictors.

In our case, ACE cannot be directly applied. So, a genetic algorithm[20] has been used to select a subset of predictors with the best predictive ability.

Principal component analysis (PCA) and cluster analysis (CA) have been used to help in the interpretation of the results.

PCA was performed on the autoscaled data, to avoid the effect of the different scale of the predictors (the ranges are between 0.06 of 2′-ribose and 280,000 of piID). Table 2 shows how the first component explains a very large fraction of the variance (66%). The second and the third components explain a significant fraction of the variance, then the eigenvalues decrease almost regularly that which usually indicates that they do not carry interesting information.

The Biplot in Figure 6 indicates that the Response has a very small loading on the first component, so that the predictors with large loadings on this component and small loading on the other components are non-informative.

In CA, the similarity between two variables (predictors and Response) has been described by the absolute value of their correlation coefficient. The matrix of the similarity can be used to single out clusters of similar variables by means of the agglomeration technique 'single linkage unweighted.'[10] The dendrogram was modified by means of 'seriation,'[19] a procedure that swaps the predictors in the similarity matrix to obtain a better order of their position in the abscissa of the dendrogram of similarities. Part of the so-obtained histogram is shown in Figure 7. In the dendrogram, only the two extremes are represented. In the upper part, it is possible to recognize the Response, so that the upper part of the dendrogram contains the predictors with a generally high correlation with the Response. Instead, the lower part of the dendrogram contains the predictors with a generally poor correlation to the variables in the upper part.

3. Results and discussion

The most important results of the regression techniques are listed in Table 3. The results obtained with GOLPE are given in a separate table (Table 4) because of the multiple possibility of selection for a given X-block. The eight techniques can be divided into three groups: (a) PLS, working on the complete matrix of the predictors; (b) SOLS, IPWPLS, ISEPLS, UVEPLS, and GOLPE that perform, with a different criterion, the selection of the relevant predictors; (c) QSR and ACE that are nonlinear techniques working with a selection of predictors.

The results of all the techniques more or less depend on the value of the following selectable parameters:

- the criterion used to select the optimum complexity of the model in PLS and related methods;
- the number of CV groups in PLS and related methods;
- the critical value of F-to-enter and of F-to-remove in SOLS;
- the number of predictors cancelled in each cycle in ISEPLS;
- the value of α, the number of noisy artificial predictors, and the randomization seed in UVEPLS;
- the number of reduced models and of dummy variables in GOLPE.

Therefore, results are not very indicative of the 'goodness' of a selection technique even if the technique is correctly applied. Moreover, a regression model must be evaluated not only with reference to the predictive ability, but also on the basis of the number of predictors

Table 4. Results obtained with GOLPE

X-Block	A		B		C		D		E		F		G		H		I		L		M	
Elimination level	Predictors	% Prediction R^2 CV	Predictors	% Prediction R^2 CV	Predictors	% Prediction R^2 CV	Predictors	% Prediction R^2 CV	Predictors	% Prediction R^2 CV	Predictors	% Prediction R^2 CV	Predictors	% Prediction R^2 CV	Predictors	% Prediction R^2 CV	Predictors	% Prediction R^2 CV	Predictors	% Prediction R^2 CV	Predictors	% Prediction R^2 CV
None	332	47.71	173	38.99	112	39.60	76	46.55	60	53.18	55	47.58	50	49.92	43	50.00	40	49.83	33	51.71	28	33.78
I	317	56.58	156	57.94	103	57.21	69	53.60	51	60.83	46	58.26	43	51.11	32	81.71	31	86.78	25	81.20	23	55.35
II	167	78.38	82	70.55	54	73.89	37	68.54	30	64.38	29	75.28	25	62.82	21	80.19	**17**	**88.16**	21	85.64	13	78.87
III	21	48.37	21	63.53	14	35.84	14	59.29	11	58.32	14	68.48	10	62.34	**11**	**87.74**	8	72.02	10	77.21	8	62.89

Table 5. Predictor selected by SOLS

A	B	C	D, E, F	G, H, I, L	M
E binding	E binding	E binding	E-binding	E-binding	E-binding
HATSm [a]	C6	4′-Ribose	4′-Ribose	2′-Ribose	2′-Ribose
RDF070v	4′-Ribose	RBN	NCIC	TIE	T(N··Cl)
RDF085u	*O*-Ribose	Ram	HATSm	H0m	TE2
Mor04u	D/Dr06	SRW05			
RDF025u	SP02	RDF060m			
L2m	RDF120m	HATSm			
	HATSm	RTu			
	*n*CaH	*n*HDon			

The gray background indicates predictors that are eliminated in the next X-block(s).

[a] HATSm removed by F-to-remove F test after L2m entered.

selected and of its complexity (the number of latent variables in the case of PLS).

However, some general trends can be observed and, finally, the objective of the selection is to have a reliable evaluation of the most important predictors and of the validity of the associated regression models (Tables 3 and 4).

With PLS, the predictors fit the Response almost perfectly, but the predictive ability is very poor. The explained variance increases from 40% to about 50% with the decreasing number of predictors (from X-block A to X-block L), if the elimination based on the correlation coefficient is not too heavy, as in X-block M. The elimination of very correlated predictors seems to have a positive effect, although it is not sufficient to obtain a model with good predictive performances.

Other linear biased techniques generally behave better than PLS. SOLS selects a very small number of predictors and shows a good predictive ability, while IPWPLS selects more predictors but with a significant improvement of the predictive importance in comparison with SOLS. ISEPLS chooses about a double amount of the predictors selected by SOLS and IPWPLS, but with the highest predictive ability. However, the increase of predictive ability from IPWPLS to ISEPLS seems not so large to justify the increase in the number of predictors. GOLPE produces good results when the preliminary selection has been rather important so that it works on only 43 variables (X-block H) or less. The results obtained with UVEPLS are rather irregular. Contrary to GOLPE, UVEPLS behaves better when the number of original predictors is rather large.

It must be noticed that IPWPLS not only selects but also weighs predictors. On the contrary, ISEPLS in the PLS algorithm uses the autoscaled predictors without further weighting. Suitable weights applied to the predictors can improve the prediction ability and generally the weights applied by IPWPLS decrease SEP in a significant way. Moreover, as concerns UVEPLS it should be underlined that it does not use the prediction ability to select the predictors. In ISEPLS, IPWPLS, and GOLPE techniques, the evaluation set is more a set of predictive optimization than a true evaluation set. So, it can be concluded that these techniques are optimistic and UVEPLS 'honest.'

A further element that should be evaluated is that SOLS and IPWPLS do not select very correlated predictors. When two or more predictors are useful but bring almost the same information, SOLS always and IPWPLS generally select only one of them. On the contrary, both UVEPLS and GOLPE select all the similar useful predictors.

Moreover, it is necessary to recognize that with the above-discussed techniques, based on the selection of useful predictors, a large predictive ability can be random when the number of objects is low and the number of predictors is very large. The effect of chance can be lower than that foreseen because of the correlation between the predictors, so that we do not have 332 or 173 or 112 or less independent predictors but as many as the number of significant components of the X-blocks (18, 16, and 14, respectively, for X-blocks A–C, according to the Kaiser criterion[24]).

A further encouraging feature is that the different regression techniques applied in this study produce similar results also in the selection of the predictors, as shown in Tables 5–7. In fact, many predictors are selected among those ones in the upper part of the dendrogram in Figure 7 (high correlation with the Response), some among those ones in the lower part (low correlation with the Response), and few of them in selected regions not shown in Figure 7. This means that different approaches converge on selecting the same predictor or, in any case, very similar predictors, generally from the same cluster.

The selection of predictors with small correlation with the Response seems a nonsense. Referring to the case of SOLS, after the first predictor has been selected, the second predictor will be the one most correlated with the residuals. Obviously, the residuals and the first predictor are orthogonal (no correlation between them), so that the second predictor will preferably be orthogonal, with very small correlation with the first one.

Figure 8 and Table 8 show the predictors selected by SOLS, IPWPLS, UVEPLS, and GOLPE with reference to the dendrogram in Figure 7.

Table 6. Predictors selected by IPWPLS, with their weight

A, B[a]		C		D		E		F, G[a]		H		I, L[a]		M	
Variable	Weight	Variable	Weight	Variable	Weight	Variable	Weight	Variable	Weight	Variable	Weight	Variable	Weight	Variable	Weight
E binding	0.110533	E binding	0.095121	E binding	0.089708	E binding	0.086869	E binding	0.123884	E binding	0.087201	E binding	0.089465	E binding	0.329658
C6	0.168778	C6	0.045941	N6	0.085405	C6	0.000000	N6	0.084999	C6	0.136242	C6	0.142211	2′-Ribose	0.090472
4′-Ribose	0.087295	N6	0.051492	2′-Ribose	0.062060	N6	0.084936	2′-Ribose	0.081830	4′-Ribose	0.164908	4′-Ribose	0.159024	4′-Ribose	0.195265
DELS	0.112098	2′-Ribose	0.037000	4′-Ribose	0.040533	*n*DB	0.068206	4′-Ribose	0.083952	SPI	0.172910	SPI	0.177862	TE2	0.197576
RDF030m	0.165517	4′-Ribose	0.050556	*n*DB	0.079042	T(N··N)	0.068873	*n*DB	0.093647	T(O··O)	0.103327	T(O··O)	0.101614	*n*CrHR	0.187029
RTu	0.171681	*n*DB	0.082616	DELS	0.047321	RDF045u	0.290725	SEigp	0.046947	RDF060m	0.051634	RDF060m	0.081783		
*n*CaH	0.184099	SP04	0.075777	RDF045u	0.109260	RDF030m	0.105222	RDF030m	0.092484	L2m	0.034548	*n*CrH2	0.248040		
		RDF045u	0.214682	RDF030m	0.092640	RTu	0.295170	RDF050m	0.084577	*n*CrH2	0.249215				
		RDF030m	0.076305	RDF050m	0.055277			RDF060m	0.072155	H-046	0.000014				
		RTu	0.191036	RTu	0.170777			RDF070m	0.041496						
		*n*CrH2	0.050873	RTv	0.098335			Rtu	0.057413						
		*n*HAcc	0.028602	*n*CrH2	0.069643			*n*CrH2	0.080100						
								*n*CrHR	0.056516						

[a] The same predictors have been selected, with about the same weights (the weights in the table refer to X-blocks A, F, and I).

E binding, the predictor with the largest correlation coefficient with the Response, is always selected. In addition, among the predictors with high correlation with the Response, C6, 2′-ribose, 4′-ribose, RDF060m, *n*CrH2, H-046, and RTu are very frequently retained.

Among the predictors with the minimum correlation with the Response, the selection techniques usually chose at least two lowest in the family HATSm, RDF030m, SPI and two or three predictors in the family *n*CrHR, BAC, *n*DB, T(O··O).

So the following predictors can be considered the set of surely informative predictors: E binding, C6, 2′-ribose, 4′-ribose, HATSm, RDF030m, SPI, *n*CrHR, BAC, RTu, T(O··O).

The PLS model derived from these 11 predictors has a leave-one-out explained variance of about 70% with only two latent variables: it is economical (11 predictors), simple (2 latent variables), and its predictive ability, compared with the extreme values of 95% or more reached by ISEPLS, IPWPLS, and SOLS, can be considered reasonable.

The loadings of the two latent variables are shown in Figure 9A. The first PLS component is clearly associated to the predictors with a large correlation to the Response, the second component is associated to the predictors with a small correlation to the Response. The second PLS component is mainly used to improve the prediction of objects **2** and **13** (Figure 9B), much less that of objects **3**, **14**, and **21**. The prediction error of the other objects is almost the same as with one or two latent variables.

Nonlinear techniques are more sensitive than linear ones to chance correlation. So their results can be accepted only in the case of very clear relationships. In spite of the apparently excellent prediction power, the results obtained with ACE are not so important. Frequently ACE confirms a linear relationship between Response and predictor. A good model has never been obtained just with 1–3 predictors.

The effect of chance is important in Quadratic stepwise selection. The selection of a new predictor continuously decreases the leave-one-out prediction error. The number of selections for data in Table 3 has been obtained by a procedure similar to the Scree test[10,11,19] shown in Figure 10 for X-Block B. After the fifth predictor is entered, the CV explained variance increases almost linearly, and it seems that this fact indicates casual correlation, so only the first five predictors selected can be considered as significant. However, in spite of this test, the results of nonlinear regression can be considered positive because they generally confirm the choice of the predictors selected by the linear methods, so that we can conclude that there is not any clear nonlinear relationship with a small number of predictors.

The results obtained with the 18 molecules of the reduced calibration set are less homogeneous than those

Table 7. Predictors selected by UVEPLS

A	B	C	D	E	F	G	H	I	L	M
E binding	E binding	E binding	E binding	E binding	E binding	E binding	E binding	E binding	E binding	E binding
C6	C6	C6	C6	C6	C6	C6	C6	C6	RDF060m	C6
C2	C2	C2	*n*BM	*n*BM	RBN	RDF030m	RDF060m	RDF060m	*n*CrH2	*n*DB
R2	*n*DB	*n*BM	*n*DB	RDF045u	*n*DB	RDF050m	*n*CrH2	*n*CrH2		T(O··O)
*n*BM	*n*AB	BAC	RDF045u	RDF030m	TIE	RDF060m	H-046	NCrHR		
*n*DB	SMTI	VRA1	RDF030m	RDF060m	RDF030m	RDF070m		H-046		
*n*AB	BAC	RDF045u	RDF060m	*n*CrH2	RDF060m	*n*CrH2				
SMTI	TE2	RDF070u	RDF070m		RDF070m	H-046				
TI1	RDF045u	RDF030m	RTu		*n*CrH2					
TI2	RDF060u	RDF050m	RTe		H-046					
D/D	RDF070u	RDF060m	*n*CrH2PLS							
BAC	RDF050m	RDF070m								
Eig1Z	RDF060m	RDF030v								
Eig1v	RDF070m	RTu								
TE1	RDF030v	*n*CrH2								
TE2	L2m	H-046								
HOMT	RTu									
DP09	*n*CrH2									
DP12	*n*CaH									
SP07	C-024									
SP09										
SP10										
SP11										
RDF035u										
RDF070u										
RDF050m										
RDF070v										
RDF060p										
RDF070p										
Am										
Av										
Vm										
HTv										
HTp										
RTu										
*n*CaH										

obtained with total set because the loss of four molecules is important. However, it is possible to find some fairly good models, generally built on intermediate data set, where the more correlated predictors have been already deleted.

Once more PLS regression gives models with very good fitting ability but poor prediction, the best result having been obtained on data set E. Other techniques based on PLS algorithm, as ISE, IPW, UVE or GOLPE, can give best results choosing accurately setting parameters. Table 9 shows, as an example, some obtained results.

More important considerations can be made about the predictors selected by all the applied techniques since all the 11 predictors underlined as important in the first part of this study are included between the first variables selected by each regression method.

PLS regression applied considering only these 11 variables gives results similar to those obtained on the complete set, without relevant variation in the *b* model coefficients. Model explains, with two latent variables 82% of total variance in fitting, 64% in leave-one-out cross-validation, and 84% of total variance of the external test set.

Thus, the results of our chemometric approach converge in identifying 11 predictors, selected both by linear and nonlinear methods, which are extremely important in determining a good affinity of the ligand toward the receptor. The E binding predictor, namely the receptor–agonist binding energy calculated with the docking program QXP,[4,25] is the descriptor which displays the highest correlation with the biological Response. Its important role could be deduced from the fact that this predictor should encode more clearly than other ones the enthalpy of binding and thus for the affinity. However, as it has widely been discussed in the literature,[26] many different docking programs are not able to correctly evaluate the various terms contributing to the quantification of the free energy of the binding process, due to the 'goodness,' of their scoring functions. In this study, the chemometric results confirm the ability of the docking program QXP to correctly evaluate binding free energies, in order to rank different ligand–receptor complexes. Moreover, they give further evidence to the reliability of the theoretical model of the human

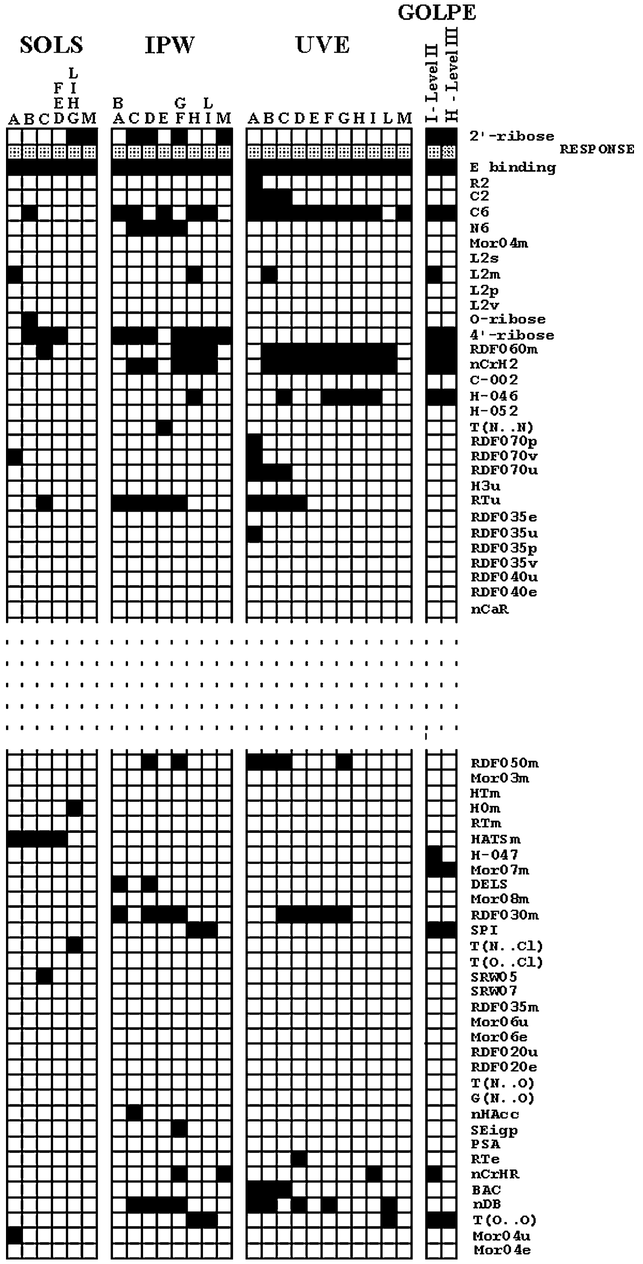
GOLPE
SOLS
IPW
UVE
L
F I
E H
A B C D G M
B G L
A C D E F H I M
A B C D E F G H I L M
I - Level II
H - Level III
2'-ribose
RESPONSE
E binding
R2
C2
C6
N6
Mor04m
L2s
L2m
L2p
L2v
O-ribose
4'-ribose
RDF060m
nCrH2
C-002
H-046
H-052
T(N..N)
RDF070p
RDF070v
RDF070u
H3u
RTu
RDF035e
RDF035u
RDF035p
RDF035v
RDF040u
RDF040e
nCaR
RDF050m
Mor03m
HTm
H0m
RTm
HATSm
H-047
Mor07m
DELS
Mor08m
RDF030m
SPI
T(N..Cl)
T(O..Cl)
SRW05
SRW07
RDF035m
Mor06u
Mor06e
RDF020u
RDF020e
T(N..O)
G(N..O)
nHAcc
SEigp
PSA
RTe
nCrHR
BAC
nDB
T(O..O)
Mor04u
Mor04e

Figure 8. Scheme of different selections of predictors with different techniques. Black boxes correspond to the selected predictors ordered as in the dendrogram of similarities (Fig. 7).

A_1 receptor recently published by some of us.[4] Besides, the descriptor C6, atomic charge on carbon atom on position 6 in the adenosine ring, and the descriptor N6, charge on position N6, are highlighted as important parameters. This result is in agreement with previous SAR studies[5a] and with our recent finding,[4] and underlines the importance of specific substituents at position 6 of the purine core for determining a high affinity toward A_1 receptor. The descriptors 2′-ribose and 4′-ribose, namely the atomic charge on positions 2′ and 4′ of the ribose ring, according to our recent docking studies[4] were expected to be related to the affinity toward A_1 adenosine receptor. However, docking studies did not allow estimating their real importance in comparison with the influence exerted by modifications of electronic features on other key positions of the adenosine structure. This information is now 'extractable' from the dendrogram of Figure 7, which underlines that electronic properties on the positions 2′ and 4′ clearly play an important role in determining the A_1 affinity value. In this study, GETAWAY (GEometry, Topology, and Atoms-Weighted AssemblY) descriptors,[27,28] implemented in DRAGON, were able to decode a large part of the information deriving from the A_1 AR–ligand interaction. They are based on a leverage matrix, similar to the one defined in statistics and employed for regression diagnostics, called molecular influence matrix (MIM). This MIM is used as molecular representation calculated from the spatial coordinates of the molecule atoms in a chosen conformation. The derived descriptors try to match 3D-molecular geometry, provided by the MIM and atom relatedness by molecular topology, with chemical information by using different atomic weightings (e.g., atomic mass, polarizability, van der Waals volume, and electronegativity). GETAWAY descriptors are divided into two different sets: H-GETAWAY, derived by using only the information provided by the MIM and R-GETAWAY, derived by joining the MIM information with the geometric interatomic distances in the molecule. In this study, the HATSm (leverage-weighted total index/weighed by atomic masses) and RTu (R total index/unweighted) descriptors became the most predictive ones. Since these variables take into account different important contributes, the information deduced from their correlation with the A_1 affinity value is not as neat as in the case of E binding. In any case, these results underline that 3D-information, derived from the conformations selected by our docking studies,[4] is essential for describing in a complete way the binding event. In addition, radial distribution function (RDF) descriptors,[29] implemented in DRAGON, and encoding for 3D-structure information, became useful in describing the same event. Interestingly, a very recent study published by Gonzáles[30] on A_{2B} agonists has identified RDF descriptors as important for deriving a useful QSAR model. Taking into account the high degree of similarity existing among adenosine receptors and among adenosine agonists, the agreement of our results with those from Gonzáles states, once more, the role of RDF descriptors and confirms the validity of the chemometric results obtained. Moreover, for an optimal predictive ability of the above-discussed regression models, some topological descriptors, namely superpendentic index (SPI), balaban centric index (BAC), and T(O··O) (sum of topological distances between O··O), plus some functional and constitutional descriptors, namely *n*CrHR (number of ring tertiary C (sp3)) and *n*DB (number of double bonds), are required. They can all be included in the large family of 2D descriptors, extensively used in QSAR and in the screening of virtual libraries, since they are rapidly computable from the molecular formula, without doing any experiment. 2D descriptors do not consider information deriving from conformational aspects but encode very important

Table 8. Definition and description of the most selected variables in SOLS, IPWPLS, UVEPLS, and GOLPE regression models

Symbol	Definition	Class
RBN	Number of rotatable bonds	Constitutional descriptors
*n*DB	Number of double bonds	Constitutional descriptors
Ram	Ramification index	Topological descriptors
SPI	Superpendentic index	Topological descriptors
D/Dr06	Distance/detour ring index of order 6	Topological descriptors
T(O··O)	Sum of topological distances between O··O	Topological descriptors
BAC	Balaban centrix index	Topological descriptors
SRW05	Self-returning walk count of order 05	Molecular walk counts
SP02	Shape profile No. 02	Randic molecular profiles
SP04	Shape profile No. 04	Randic molecular profiles
DELS	Molecular electrotopological variation	Geometrical descriptors
RDF025u	Radial Distribution Function—2,5/unweighted	RDF descriptors
RDF045u	Radial Distribution Function—4,5/unweighted	RDF descriptors
RDF085u	Radial Distribution Function—8,5/unweighted	RDF descriptors
RDF030m	Radial Distribution Function—3,0/weighted by atomic masses	RDF descriptors
RDF060m	Radial Distribution Function—6,0/weighted by atomic masses	RDF descriptors
RDF070m	Radial Distribution Function—7,0/weighted by atomic masses	RDF descriptors
RDF120m	Radial Distribution Function—12,0/weighted by atomic masses	RDF descriptors
RDF070v	Radial Distribution Function—7,0/weighted by atomic van der Waals volumes	RDF descriptors
RDF070p	Radial Distribution Function—7,0/weighted by atomic polarizabilities	RDF descriptors
Mor04u	3D-MoRSE—signal 04/unweighted	3D-MoRSE descriptors
L2m	2nd component size directional WHIM index/weighted by atomic masses	WHIM descriptors
HATSm	Leverage-weighted total index/weighted by atomic masses	GETAWAY descriptors
RTu	R total index/unweighted	GETAWAY descriptors
*n*CrH2	Number of ring secondary C(sp3)	Functional groups
*n*CrHR	Number of ring tertiary C(sp3)	Functional groups
*n*CaH	Number of unsubstituted aromatic C(sp2)	Functional groups
*n*HDon	Number of donor atoms for H-bonds (with N and O)	Functional groups
*n*HAcc	Number of acceptor atoms for H-bonds (N, O, F)	Functional groups
H-046	H attached to C0(sp3) no X attached to next C	Atom-centred fragments
C6	Atomic charge on position 6 of the adenosine ring	
N6	Atomic charge on N6 position of the adenosine ring	
2′-Ribose	Atomic charge on position 2′ (ribose) of the adenosine ring	
4′-Ribose	Atomic charge on position 4′ (ribose) of the adenosine ring	
E binding	Calculated binding energy	

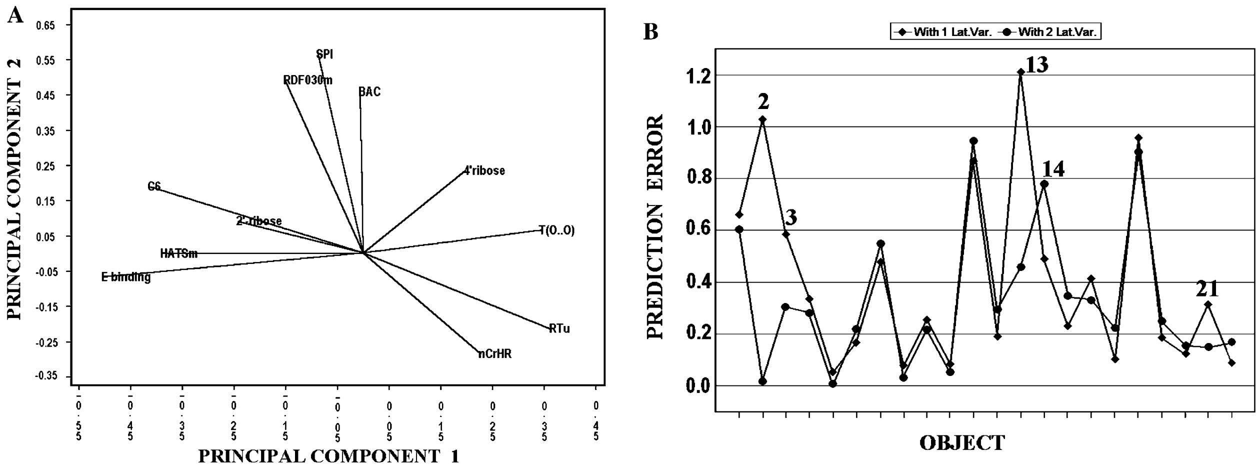

Figure 9. (A) Loadings of PLS latent variables for the model with the selected 11 predictors. (B) Absolute error of prediction with one and two latent variables for the model with the selected 11 predictors.

information on adjacency of atoms and groups, relative distances among different functional moieties, etc. Their selection in our regression models clearly demonstrates that adenosine agonists should be molecules strictly related to the scaffold of adenosine and suggests that in the case of adenosine agonists both the 3D information deriving from a putative 'active' conformation assumed by the molecule and the 2D information deriving from a high structural similarity between the agonist and the natural ligand are necessary for obtaining a good predictive ability of the QSAR model.

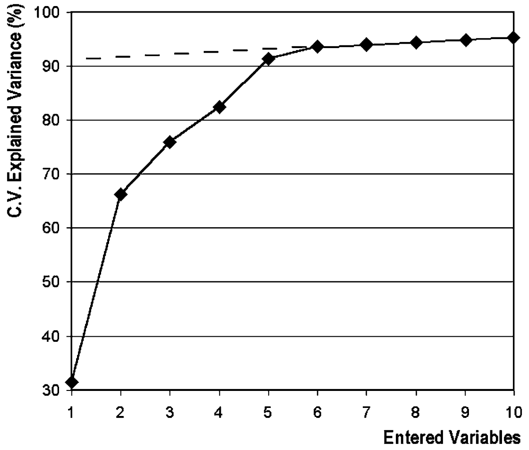

Figure 10. CV % explained variance as a function of the number of selected predictors (QSRdratic Stepwise Regression, X-Block B).

4. Conclusions

In this study, eight different prediction methods were applied to investigate the complex relationships between the K_i (nM) affinity values and the molecular structure on a data set of 21 selective A_1 adenosine agonists, plus adenosine, to find out, among a large number of descriptors encoding for the widest information, which molecular properties are more strictly related to the experimentally determined affinity values (K_i). The obtained results were carefully checked in order to avoid any by-chance correlation and are in agreement with recent literature data. They suggest that the ligand–A_1 receptor binding is mainly described by the binding energy calculated with suitable docking programs, by the electronic properties on positions C6, N6, 2′, and 4′ of the adenosine scaffold and by some 3D and 2D descriptors. The 3D and 2D descriptors encode, respectively, for information deriving from a probable 'active' conformation of the ligand and for the degree of molecular similarity between the agonist and the adenosine.

In conclusion, the results of the present work demonstrate, through a QSAR study, that affinity prediction toward A_1 AR is mainly due to specific substituents on positions N6, 2′, and 4′, which play an important role in determining the 'active' conformation and a number of positive interactions with the counterpart, and to the molecular similarity of the agonist with adenosine core, in agreement with previous SAR studies.[5a]

In addition, the chemometric approach here presented underlines the usefulness of an 'a priori' evaluation of the theoretical binding energy, through a docking study, as a screening test in the quest for new selective and highly effective compounds.

Acknowledgment

This work was financially supported by MIUR (cofin.2002, prot. 2002038577_002).

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2005.09.058.

References and notes

1. Fredholm, B. B.; Ijzerman, A. P.; Jacobson, K. A.; Klotz, K. N.; Linden, J. *Pharmacol. Rev.* **2001**, *53*, 527–552.
2. Ribeiro, J. A.; Sebastião, A. M.; de Mendonça, A. *Prog. Neurobiol.* **2002**, *68*, 377–392.
3. Bondavalli, F.; Botta, M.; Bruno, O.; Ciacci, A.; Corelli, F.; Fossa, P.; Lucacchini, A.; Manetti, F.; Martini, C.; Menozzi, G.; Mosti, L.; Ranise, A.; Schenone, S.; Tafi, A.; Trincavelli, M. L. *J. Med. Chem.* **2002**, *45*, 4875–4887.
4. Giordanetto, F.; Fossa, P.; Menozzi, G.; Schenone, S.; Bondavalli, F.; Ranise, A.; Mosti, L. *J. Comput. Aided Mol. Des.* **2003**, *17*, 39–51.
5. (a) Muller, C. E. *Curr. Med. Chem.* **2000**, *7*, 1269–1288; (b) Knutsen, L. J. S.; Lau, J.; Petersen, H.; Thomsen, C.; Weis, J. U.; Shalmi, M.; Judge, M. E.; Hansen, A. J.; Sheardown, M. J. *J. Med. Chem.* **1999**, *42*, 3463–3477; (c) Spada, A.; Fink, C. A.; Myers, M. R. U.S.P 5,364,862; 1994.
6. DRAGON 2.1. <http://www.talete.mi.it>, Todeschini, R; Consonni, V; Mauri, A; Pavan, M. Milano Chemometrics and QSAR Research Group.
7. Forina, M; Lanteri, S; Armanino, C. Q-PARVUS, An Extendable Package of Programs for Explorative Data Analysis, Classification and Regression Analysis, Dip. Chimica e Tecnologie Farmaceutiche, University of

Table 9. Results obtained on the reduced calibration set (18 objects)

Block	Regression method	No. of predictors	Complexity	SEC	% Fitting R^2	SEP	% Prediction R^2 CV	Res. Std ext	% Prediction Expt. set
E	PLS	60	10	0.006	99.99	0.567	50.30	0.332	64.15
D	ISE	8	5	0.223	91.87	0.314	83.47	0.182	89.25
E	ISE	21	5	0.098	98.41	0.295	85.42	0.247	80.04
C	IPW	10	3	0.193	93.89	0.263	89.29	0.317	67.31
D	IPW	6	3	0.202	93.33	0.265	89.11	0.293	72.05
E	IPW	6	5	0.214	92.53	0.268	88.89	0.306	69.52
A	UVE95%	41	3	0.268	88.24	0.426	71.88	0.288	72.88
F	UVE95%	8	2	0.407	72.85	0.527	57.08	0.346	61.06
G	UVE95%	9	2	0.360	78.74	0.486	63.51	0.226	83.32
D	V-GOLPE	38	10	0.009	99.99	0.423	69.92	0.328	64.95
F	V-GOLPE	26	3	0.309	84.33	0.480	61.27	0.306	69.48

Genova, Italy Available (free, with manual and examples) at <http://www.parvus.unige.it>.

8. Forina, M; Lanteri, S; Armanino, C; Cerrato Oliveros, C; Casolino, C. V-PARVUS Release 1.0, An Extendable Package of Programs for Explorative Data Analysis, Classification and Regression Analysis, Dip. Chimica e Tecnologie Farmaceutiche, University of Genova, Italy. Available free, with manual and examples, at <http://www.parvus.unige.it>, 2003.
9. Spartan ′O2, <http://www.wavefun.com>.
10. Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics*; Elsevier: Amsterdam, 1998.
11. Frank, I. E.; Todeschini, R. *The Data Analysis Handbook*; Elsevier: Amsterdam, 1994.
12. Forina, M.; Casolino, C.; Pizarro Millan, C. *J. Chemometr.* **1999**, *13*, 165–184.
13. Boggia, R.; Forina, M.; Fossa, P.; Mosti, L. *Quant. Struct.-Act. Relat.* **1997**, *16*, 201–213.
14. Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M.; Sterna, C. *Anal. Chem.* **1996**, *68*, 3851–3858.
15. Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20.
16. http://www.miasrl.com/software/golpe/manual/background.html.
17. Breiman, L.; Friedman, J. H. *J. Am. Stat. Assoc.* **1985**, *80*, 580–619.
18. Massart, D. L.; Kaufman, L. *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*; Wiley & Sons: NewYork (NY), 1983.
19. Wishart, D. *Comput. Sci. Stat.* **1998**, *29*, 48–51.
20. Lucasius, C. B.; Kateman, G. *TRAC-Trend Anal. Chem.* **1991**, *10*, 254–281.
21. Leardi, R.; Boggia, R.; Terrile, M. *J. Chemometr.* **1992**, *6*, 267–281.
22. Ojelund, H.; Madsen, H.; Thyregod, P. *J. Chemometr.* **2001**, *15*, 497–509.
23. Kennard, R. W.; Stone, L. A. *Technometrics* **1969**, *11*, 137–148.
24. Kaiser, H. F. *Educ. Psychol. Meas.* **1960**, *20*, 141–151.
25. McMartin, C.; Bohacek, R. *J. Comput. Aided Mol. Des.* **1997**, *11*, 333–344.
26. Tame, J. R. H. *J. Comput. Aided Mol. Des.* **1999**, *131*, 99–108.
27. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682–692.
28. Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693–705.
29. Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *Vib. Spectrosc.* **1999**, *19*, 151–164.
30. González, M. P.; Terán, C.; Fall, Y.; Teijeira; Besada, P. *Bioorg. Med. Chem.* **2005**, *13*, 601–608.