Journal of Computer-Aided Molecular Design **18:** 495–509, 2004.
DOI 10.1007/s10822-004-5203-7

Variable selection and specification of robust QSAR models from multicollinear data: arylpiperazinyl derivatives with affinity and selectivity for α_2-adrenoceptors

D.W. Salt[a], L. Maccari[a,b], M. Botta[b] & M.G. Ford[a,*]
[a]*The Centre for Molecular Design, IBBS, University of Porthsmouth, Porthsmouth P01 2DY, UK;* [b]*Dipartimento Farmaco Chimico Tecnologico, Universita' degli Studi di Siena, Via Aldo Moro, I-53100 Siena, Italy*

Received 18 May 2004; accepted in revised form 5 October 2004

Key words: α_2-adrenoceptors, affinity, arylpiperazinyl derivatives, canonical correlation, continuum regression, PARAGON, QSAR, selectivity

Summary

Two QSAR models have been identified that predict the affinity and selectivity of arylpiperazinyl derivatives for α_1 and α_2 adrenoceptors (ARs). The models have been specified and validated using 108 compounds whose structures and inhibition constants (K_i) are available in the literature [Barbaro et al., J. Med. Chem., 44 (2001) 2118; Betti et al., J. Med. Chem., 45 (2002) 3603; Barbaro et al., Bioorg. Med. Chem., 10 (2002) 361; Betti et al., J. Med. Chem., 46 (2003) 3555]. One hundred and forty-seven predictors have been calculated using the Cerius 2 software available from Accelrys. This set of variables exhibited redundancy and severe multicollinearity, which had to be identified and removed as appropriate in order to obtain robust regression models free of inflated errors for the β estimates – so-called bouncing βs. Those predictors that contained information relevant to the α_2 response were identified on the basis of their pairwise linear correlations with affinity ($-\log K_i$) for α_2 adrenoceptors; the remaining variables were discarded. Subsequent variable selection made use of Factor Analysis (FA) and Unsupervised Variable Selection (UzFS). The data was divided into test and training sets using cluster analysis. These two sets were characterised by similar and consistent distributions of compounds in a high dimensional, but relevant predictor space. Multiple regression was then used to determine a subset of predictors from which to determine QSAR models for affinity to α_2-ARs. Two multivariate procedures, Continuum Regression (the Portsmouth formulation) and Canonical Correlation Analysis (CCA), have been used to specify models for affinity and selectivity, respectively. Reasonable predictions were obtained using these *in silico* screening tools.

Abbreviations: AR – adrenoceptors; CCA – Canonical Correlation Analysis; cnv*f* – canonical variate first set; cnv*s* – canonical variate second set; C_i – the *i*th component; CR – Continuum Regression; GPCR – G-protein coupled receptors; $-\log K_i$ – log transformed affinity; ^{3}H – tritiated; UFS – Unsupervised Forward Selection.

Introduction

α_1- and α_2-adrenoceptors (α_1- and α_2-AR) belong to the family of the G-protein coupled receptors (GPCRs), characterised by seven α- helical domains traversing the cell membrane. Pharmacological and binding studies have demonstrated that α-ARs are comprised of multiple subtypes. α_1-ARs are classified into three subtypes, α_{1A}, α_{1B} and α_{1D} [1,2]; α_2-ARs have been classified into four subtypes, named α_{2A}–α_{2D}, respectively [3].

α_1-AR antagonists interact with the postsynaptic α-adrenoceptors, inhibiting the cascade of events usually mediated by the biogenic neurotransmitter

*To whom correspondence should be addressed. E-mail: martyn.ford@port.ac.uk

noradrenaline. In recent years, the search for new selective inhibitors of these targets has intensified because of their possible use in the treatment of hypertension and benign prostatic hyperplasia [4, 5].

Presynaptic α_2-adrenoceptors modulate the release of noradrenaline from nerve endings through a local feedback mechanism [6]. α_2-AR antagonists should therefore increase synaptic concentration of noradrenaline and the resultant postsynaptic stimulation of α- and β-ARs. Such agents may have clinical utility in the treatment of diabetes, depression, and male sexual dysfunction [7, 8]. Molecular cloning studies [9, 10] have shown that α_1- and α_2-ARs share many common features. A synthetic compound with affinity toward α-adrenergic receptors may bind to both α_1- and α_2-ARs so that the design of a drug molecule with high affinity and selectivity toward one of the α-AR is difficult.

Our first target was the design of new and potentially selective α_1-AR antagonists. In the course of our studies in this field, we have synthesized a series of compounds mainly characterized by an arylpiperazinyl moiety linked through an alkyl chain to a wide range of hydrophobic moieties. All the compounds have been tested for their affinity for α_1- and α_2-ARs by determining for each compound its ability to displace [^{3}H]prazosin or [^{3}H]rauwolscine from specific binding sites on rat cerebral cortex. K_i values were determined on the basis of three competition-binding experiments in which seven drug concentrations, run in triplicate, were used [6, 11–13].

The purpose of this paper is to describe the development of robust QSAR models to predict α_2-AR binding affinity and selectivity for this class of receptors. The aim was to develop models that retained consistency with a previously defined pharmacophore model for α_1 adrenoceptors. Two robust QSAR models have been identified that should aid rational design of new and effective compounds with selectivity for α_1- or α_2-ARs.

Developing a robust QSAR model for predicting the affinity of arylpiperazinyl derivatives for α_2-ARs

The response and descriptor sets

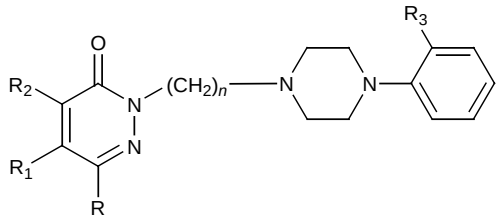
One hundred and eight arylpiperazinyl derivatives were synthesised and their measured α_1 and α_2 affinities, determined as K_i values, are given in Table 1. These values ranged from 0.10 to 986 nM for α_1 receptors and from 0.56 to 2072 nM for α_2 receptors. The affinity data were transformed into $-\log K_i$ values to give affinity estimates with normally distributed errors [6, 11–13]. All the compounds of the data set were built using the two- and three-dimensional sketcher of the software Catalyst [14]. A representative family of conformations was generated for each molecule using the poling algorithm and the best quality conformational analysis [15–17] and CHARMm force field implemented in Catalyst [18]. The poling function prevents conformations of molecules from being too close together and allows the exploration of the conformational space of molecules within a user-defined energy threshold. The best quality analysis is also considered as the method of choice for systems with flexible rings. Conformational diversity was emphasized by selection of the conformers that fell within the 20 kcal/mol range above the lowest-energy conformation that was found.

For each molecule, the lowest energy conformer was selected as the structure upon which to base the calculation of molecular properties. Each low energy structure was imported to Cerius 2 and atomic charges were assigned using the charge equilibration method. Descriptors encoding spatial, structural, electronic, thermodynamic, quantum mechanical, topological, and electro-topological properties were calculated (Table 2). This set of 147 descriptors is characterised by extreme redundancy and multicollinearity. Moreover, many of the descriptors are unrelated to the affinity and selectivity of these compounds for adrenoceptors. The task, then, is to identify a small subset that can be used to predict these biological properties and to specify robust prediction models to guide future programmes of synthesis.

Splitting the data into a training and test set based on relevant descriptors

Before splitting the data into test and training sets, it is important to eliminate any irrelevant variables, thus ensuring that model construction and validation is based on an appropriate, low dimensional property space. The set of descriptor

Table 1. Compound structures for 108 arylpiperazinyl derivatives and their α_1 and α_2 affinities for the adrenoceptor.

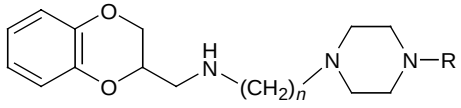


Comp.	Ref.	Name in ref.	K_i (nM)		R^b	$R_1{}^b$	$R_2{}^b$	R_3	n
			α_1-AR	α_2-AR					
1[a]	12	5a	472	2072	A	–H	–H	–Cl	2
2[a]	12	5b	48.3	436	A	–H	–H	–Cl	3
3	12	5c	24.6	65.2	A	–H	–H	–Cl	4
4	12	5d	23.3	134	A	–H	–H	–Cl	5
5	12	5e	29.7	138	A	–H	–H	–Cl	6
6	12	5f	4.6	24.5	A	–H	–H	–Cl	7
7[a]	12	2a	33	350	B	–H	–H	–OMe	2
8[a]	12	2b	6.5	158	B	–H	–H	–OMe	3
9[a]	12	2c	1.1	16	B	–H	–H	–OMe	4
10[a]	12	4a	237	1492	A	–H	–H	–OMe	2
11	12	2d	3.1	30.5	B	–H	–H	–OMe	5
12	12	2e	6.4	113	B	–H	–H	–OMe	6
13	12	2f	1.7	89.8	B	–H	–H	–OMe	7
14[a]	12	3a	70.4	265	B	–H	–H	–Cl	2
15	12	3b	15.2	44.7	B	–H	–H	–Cl	3
16	12	3c	1.3	5.8	B	–H	–H	–Cl	4
17	12	3d	4.5	4.5	B	–H	–H	–Cl	5
18[a]	12	3e	15.1	31.5	B	–H	–H	–Cl	6
19[a]	12	3f	1	10.3	B	–H	–H	–Cl	7
20[a]	12	4b	115	673	A	–H	–H	–OMe	3
21[a]	12	6a	22.7	1098	C	–H	–H	–OMe	2
22[a]	12	6b	3.2	170	C	–H	–H	–OMe	3
23[a]	12	6c	0.9	20	C	–H	–H	–OMe	4
24	12	6d	6.8	95.6	C	–H	–H	–OMe	5
25	12	6e	2.8	104	C	–H	–H	–OMe	6
26	12	6f	1.1	306	C	–H	–H	–OMe	7
27	12	4c	47.5	393	A	–H	–H	–OMe	4
28[a]	12	4d	37.5	427	A	–H	–H	–OMe	5
29	12	4e	52.5	131	A	–H	–H	–OMe	6
30[a]	12	4f	6	76.5	A	–H	–H	–OMe	7
31	11	1a	38	1261	D	–H	–H	–OMe	2
32[a]	11	1b	7.3	254	D	–H	–H	–OMe	3
33[a]	11	1c	0.6	62	D	–H	–H	–OMe	4
34[a]	11	1d	12.9	69.8	D	–H	–H	–OMe	5
35	11	1e	7	71.8	D	–H	–H	–OMe	6
36[a]	11	1f	2.3	45.8	D	–H	–H	–OMe	7
37	11	1g	18	370	D	–H	–H	–Cl	2
38	11	1h	6.8	316	D	–H	–H	–Cl	3
39	11	1i	0.8	69.4	D	–H	–H	–Cl	4
40	11	1j	7	138	D	–H	–H	–Cl	5
41	11	1k	8.4	138.8	D	–H	–H	–Cl	6
42	11	1l	15	139	D	–H	–H	–Cl	7

Table 1. Continued.

Comp.	Ref.	Name in ref.	K_i (nM)		R^b	$R_1{}^b$	$R_2{}^b$	R_3	n
			α_1-AR	α_2-AR					
43	11	4e	180	790	E	–H	–H	–OMe	4
44[a]	11	4g	43.7	255	E	–H	–H	–OMe	5
45[a]	11	4i	17.8	100	E	–H	–H	–OMe	6
46[a]	11	4k	54.7	96	E	–H	–H	–OMe	7
47	11	4d	12.8	260	E	–H	–H	–Cl	3
48	11	4f	7.9	24	E	–H	–H	–Cl	4
49[a]	11	4h	13.3	105	E	–H	–H	–Cl	5
50	11	4j	5.9	23.7	E	–H	–H	–Cl	6
51[a]	11	4l	5.6	23.5	E	–H	–H	–Cl	7
52[a]	11	2a	16	409	–H	D	–Cl	–OMe	2
53[a]	11	2j	10	39.3	–H	D	–Cl	–Cl	4
54	11	2k	4.5	29	–H	D	–Cl	–Cl	5
55[a]	11	2l	4.2	25.2	–H	D	–Cl	–Cl	6
56[a]	11	2m	2.7	7.4	–H	D	–Cl	–Cl	7
57	11	2n	5.6	22.8	–H	D	–Cl	–Cl	8
58	11	3c	33	440	–H	E	–Cl	–OMe	2
59	11	3a	20.4	680	–H	E	–Cl	–OMe	3
60[a]	11	3e	33.5	870	–H	E	–Cl	–OMe	4
61[a]	11	2c	4.3	230	–H	D	–Cl	–OMe	4
62	11	3g	1.9	23.3	–H	E	–Cl	–OMe	5
63[a]	11	3i	4.1	24.5	–H	E	–Cl	–OMe	6
64[a]	11	3k	1.9	520	–H	E	–Cl	–OMe	7
65	11	3d	3.9	150	–H	E	–Cl	–Cl	2
66[a]	11	3b	3.5	85	–H	E	–Cl	–Cl	3
67	11	3f	19	280	–H	E	–Cl	–Cl	4
68[a]	11	3h	3.6	18	–H	E	–Cl	–Cl	5
69	11	3j	5.7	66	–H	E	–Cl	–Cl	6
70[a]	11	3l	11.5	36.2	–H	E	–Cl	–Cl	7
71	11	2d	3.9	15	–H	D	–Cl	–OMe	5
72[a]	11	2f	1.4	4.6	–H	D	–Cl	–OMe	7
73	11	2g	3.5	22.7	–H	D	–Cl	–OMe	8
74[a]	11	2h	58.8	292.3	–H	D	–Cl	–Cl	2
75	11	2i	27.8	219	–H	D	–Cl	–Cl	3
76	11	2e	1.5	3.5	–H	D	–Cl	–OMe	6
77[a]	11	2b	14.5	245	–H	D	–Cl	–OMe	3
78	6	13	0.55	1.59	–H	F	–Cl	–OMe	7
79[a]	6	14	0.43	2.04	–H	F	–Cl	–OEt	7
80[a]	6	15	0.26	3.21	–H	F	–Cl	–O^{i}Pr	7
81	6	16	0.58	8.22	–H	D	–Cl	–OEt	7
82[a]	6	2	0.5	4.0	–H	E	–Cl	–OEt	7
83	6	3	0.052	0.56	–H	E	–Cl	–O^{i}Pr	7
84	6	7	1.68	0.85	–H	G	–Cl	–OMe	7
85	6	8	0.42	2.16	–H	G	–Cl	–OEt	7
86[a]	6	9	0.31	3.34	–H	G	–Cl	–O^{i}Pr	7
87	6	4	0.37	1.23	–H	–Cl	E	–OMe	7
88[a]	6	5	0.23	0.80	–H	–Cl	E	–OEt	7
89	6	6	0.08	0.66	–H	–Cl	E	–O^{i}Pr	7
90[a]	6	10	0.54	1.45	–H	–Cl	G	–OMe	7
91[a]	6	11	0.43	2.40	–H	–Cl	G	–OEt	7
92[a]	6	12	0.32	13.4	–H	–Cl	G	–O^{i}Pr	7

Table 1. Continued.



Comp.	Ref.	Name in ref.	$K_{i\ (nM)}$		*R*	*n*
			α1-*AR*	α2-*AR*		
93[a]	13	10	986	163.7	*p*-Methoxyphenyl	2
94	13	3	16.5	134	*o*-Methoxyphenyl	3
95	13	4	113.7	221.1	*o*-Chlorophenyl	2
96	13	5	127	104.9	*o*-Chlorophenyl	3
97[a]	13	7	147.7	380.2	Phenyl	3
98	13	9	346.7	25.6	2-Pyridinyl	3
99[a]	13	11	718.2	96.8	*p*-Methoxyphenyl	3
100	13	12	359.4	30.0	*o*-Fluorophenyl	2
101	13	13	203	855.4	*o*-Fluorophenyl	3
102[a]	13	15	292.5	140.8	2-Methyl-4-chloropyridazin-3(2*H*)-one-5-yl	3
103	13	17	331	61.0	2-Furoyl	3
104[a]	13	2	37	67.0	*o*-Methoxyphenyl	2
105[a]	13	6	201.3	186.7	Phenyl	2
106	13	8	415.8	238.8	2-Pyridinyl	2
107[a]	13	14	269	293.1	2-Methyl-4-chloropyridazin-3(2*H*)-one-5-yl	2
108	13	16	177.2	158.5	2-Furoyl	2

[a]Compounds belonging to the training set.
[b]A, 1-imidazolyl; B, 1-benzimidazolyl; C, 1-indolyl; D, 4-[2-(2-methoxyfenoxy)ethyl]piperazin-1-yl; E, 4-(2-furoyl)piperazin-1-yl; F, 4-[2-(2-ethoxyphenoxy)ethyl]piperazin-1-yl; G, 4-[2-(1,4-benzodioxan)]piperazin-1-yl.

variables was first investigated to identify relevant variables, those with significant correlation with $-\log K_i$, which could be used to explain the responses ($-\log K_i$) to be modelled, where K_i has units of nM. Variables with non-significant correlations with $-\log K_i$ at the 5% significance level (i.e., *P*-value less than 0.05) and with a small variation (i.e., coefficient of variation below 5%) were eliminated to give a subset of 93 relevant descriptors. Cluster analysis of this reduced descriptor set was then used to assign compounds to training and test sets to ensure that they have similar X-block characteristics (i.e., occupy the same property space). Four major clusters were identified. Random samples were taken from each cluster and pooled to obtain a training set of 53 and a test set of 55 compounds. Principal component analysis (PCA) was applied to each set to establish whether the training and test set contained molecules that spanned the descriptor space. The variances (eigenvalues) and the proportions of variances explained by the first two PCAs of the training set are 64.00 and 9.22 and 0.69 and 0.10, respectively. Those for the test set are 60.42 and 9.26 and 0.65 and 0.10, respectively. The pattern of the score plots for the first two components for the training and test sets confirmed that the distributions of the two sets were similar and therefore show the consistency required for validation of QSAR models.

Removing the redundancy and adjusting the multicollinearity of the descriptors

Analysis of the correlation matrix for the training set descriptors using the unsupervised forward selection (UFS) algorithm [19] confirmed that high levels of redundancy and multicollinearity remained. The next task was to identify a subset of the variables based on the training set that could explain the response variable ($-\log K_i\alpha_2$), and could therefore be used to construct a model. This set should contain no redundant variables and its multicollinearity structure should be identified

Table 2. Descriptor families used to specify QSAR prediction models for the affinity and selectivity of arylpiperazinyl derivatives for α_2-adrenoceptors.

Descriptor families					
Spatial	RadOfGyration	3D	Structural	MW	2D
	Jurs descriptors	3D		Rotlbonds	2D
	Shadow Indices	3D		Hbondacceptor	2D
	Area	3D		Hbonddonor	2D
	Density	3D		Chiralcenters	2D
	PMI	3D			
	Vm	3D			
Electronic	Charge	2D	E_State_Keys		2D
	Fcharge	2D			
	Apol	2D			
	Sr	2D			
Thermodynamic	AlogP98	2D	QM	LUMO_mopac	3D
	AlogP_atypes	2D		Dipole_mopac	3D
	Fh2o	2D		HF_mopac	3D
	Foct	2D		HOMO_mopac	3D
	LogP	2D			
	MR	2D			
	MolRef	2D			
Topological	Balaban	2D			
	Kappa Indices	2D			
	PHI	2D			
	SubgraphCount	2D			
	Chi indices	2D			
	Wiener	2D			
	log Z	2D			
	Zagreb	2D			

[20], so that only the population multicollinearity is retained in the prediction models. Sample multicollinearity that has arisen by chance should be removed.

Factor analysis of the reduced set of 93 descriptor variables plus the response ($-\log K_i$) was undertaken to identify clusters of highly associated, relevant properties with weak inter-property associations. Factors showing a significant ($P < 0.05$) loading for the response variable were selected and representative descriptors, those with maximum loading on each factor, retained. This process reduced the 93 descriptors to a set of 17 variables from which a multiple regression model was constructed. Six of these variables (AC2, AC9, AC27, SssCH2, SC3C and MW) made significant ($P < 0.05$) contributions to the explanation of $-\log K_i$ based on the significance of their contribution (t-test, $P < 0.05$) to a regression of $-\log K_i$ on all 17 descriptors. The UFS procedure [14] was then applied to this set of 6 descriptors. UFS is a variable selection method that eliminates redundancy and reduces multicollinearity to any specified level by removing variables associated with this feature [15]. UFS operates on the correlation matrix of the X-block and starts by identifying the two least correlated variables and adds other variables on the basis of their multiple correlation coefficient with those already chosen. UFS was applied for each case using values of R^2_{max} stepping from 0.1 to 0.9999 in increments of 0.1. All six variables contributed unique information, but formed a set containing significant multicollinearities. The challenge, therefore, was to identify a QSAR prediction model that retained the population

multicollinearity, but which was free of multicollinearities that had arisen by chance as a result of sampling. Sample multicollinearity will inflate the errors of the β estimates without contributing useful information to the QSAR model and result in unstable regression equations that cannot be used to give reliable predictions.

Model specification for predicting the affinity of compounds for the α_2-AR

Q^2_{LOO} (i.e., the model is robust, and therefore generalised). Exclusion of components C3, C4 and C6 has resulted in a contraction of the component space. The component construction criterion defined by $\alpha = 0.3758$ has sacrificed fit to identify a more robust QSAR model with good predictive power less affected by chance correlation. This is confirmed by *N*-fold cross-validation which showed a small variance for Q^2 even at $N = 2$, a test of internal prediction (Figure 2).

$$-\log K_i = 4.54MW + 1.41AC2 + 0.51AC9 - 0.48AC27 - 2.65SssCH2 - 3.21SC3C$$
$$(\pm 1.01) \quad (\pm 0.28) \quad \pm 0.17) \qquad (\pm 0.10) \qquad (\pm 0.56) \qquad (\pm 0.68) \tag{1}$$
$$n = 53; Q^2_{LOO} = 0.74; R^2 = 0.73; R^2_{adj} = 0.72; S = 0.53; F_{3,49} = 44.37; P < 0.00009$$

Regression models were built from the subsets of variables identified by UFS using the Portsmouth fixed alpha formulation of Continuum Regression – CR [21, 22]. All the calculations were performed using the PARAGON software which can be downloaded from the following website – www.port.ac.uk. The most significant model included all 6 variables. The LOO cross-validated R^2 (Q^2_{LOO}) was maximised by adjusting the CR parameter α to identify an optimum value, $\alpha = 0.3758$ (Figure 1). Only three components made a significant contribution to the explanation of $-\log K_i$ assuming one degree of freedom per component. This assumption is based on the fact that the components constructed by continuum regression are linear weighted transformations of the original variables. Three significant components (C1, C5 and C2 in order of inclusion) were selected to form a QSAR model (Equation 1) that maximised

The standard errors in parentheses for Equation 1 are estimates based on 40 bootstraps. The predictor variables have been standardised. Randomisation test tail probabilities based on 150 permutations of *y* were below 0.01 for fit and prediction. The partitioning of the DOFs for the *F*-statistic assumes a three component model for fit.

The model was used to predict the α_2 affinities of the 53 training set and the 55 test set compounds. The observed and predicted affinities for the two sets show significant correlations ($P < 0.0001$), the latter confirming that the external power of prediction was sufficiently large to be of use (Figure 3). Setting a threshold for the prediction of $-\log K_i$ at -1.5, for example, would have identified a set of 17 candidates for synthesis that would have included the 7 compounds with the highest α_2 affinities of the test set. The enrichment would have been (55/17) = 3.2-fold. Thus, the model can be used for virtual screening to predict α_2 affinity in order to identify

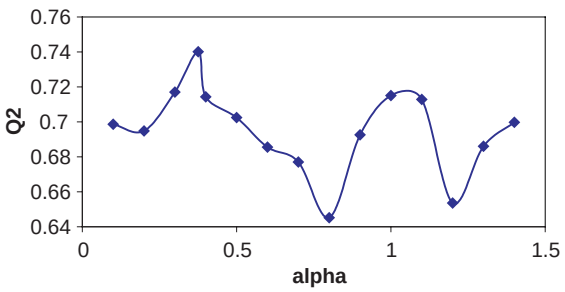


Figure 1. Effect of varying the component construction criteria α on Q^2_{LOO}.

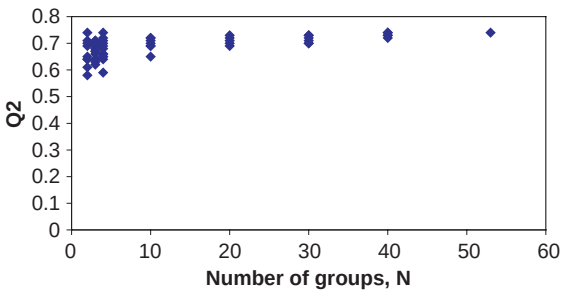


Figure 2. *N*-fold cross-validation of the 6-variable QSAR model specified using supervised and unsupervised variable selection and continuum regression.

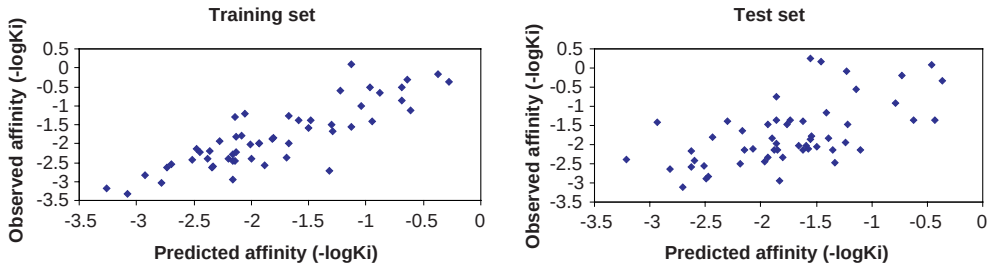


Figure 3. Plots of observed versus predicted α_2 affinities for the training and test sets.

compounds for synthesis and high throughput screening against this target receptor.

A prediction model for selectivity based on canonical correlation

Having specified a robust model with which to predict the binding affinities of the arylpiperazinyl derivatives to the α_2-AR, attention turned to the prediction of the selectivity of these compounds. As noted earlier, the arylpiperazinyl derivatives also have affinity for the α_1-AR and this could lead to unwanted side-effects unless compounds with preferential binding for the α_2-receptor can be identified. A canonical correlation analysis (CCA) was undertaken using the training set α_1 and α_2 affinity data ($-\log K_i$) to form a pair of correlated ($r = 0.71$) response (y) variables. The objective was to identify key descriptors responsible for selectivity between the two receptor types. Once again, the 17 descriptor variables identified using factor analysis provided the starting set.

Canonical correlation analysis of affinities for α_1- and α_2-ARs

CCA is a multivariate technique for analysing data where there are two distinct blocks of data with at least two variables in each block [23], e.g. data consisting of a block of response variables and a block of descriptor variables. It is customary to refer to the response variables as the Y-block and the descriptor variables as the X-block. CCA is only worth performing if the response variables are not only correlated with the X-block variables, but also with each other. By performing a simultaneous analysis on all the response variables information in their correlation structure, which is ignored when performing separate multiple regressions, is utilised.

CCA first forms a linear combination of the y-variables and a linear combination of the x-variables such that the correlation between these two linear combinations is the maximum that can be achieved amongst all such pairs. This step is followed by the identification of a second pair of variates (termed canonical variates (CVs)) which have the next highest pairwise correlation subject to them being orthogonal with both the previous variates. This process continues until $s = \min(q, p)$ pairs of CVs have been formed, where q is the number of response variables and p is the number of descriptor variables.

Denote

$$\begin{aligned} \text{cnv}f1 &= a_{11}y_1 + \cdots + a_{1q}y_q \\ \text{cnv}s1 &= b_{11}x_1 + \cdots + b_1px_p \end{aligned} \tag{2}$$

to be the first pair of canonical variates where the '*f*' and '*s*' in cnv*f*1 and cnv*s*1 refer respectively to the first and second sets of variates. The method of estimating the *a*'s and *b*'s in Equation 1 is based on an eigenvector eigenvalue analysis of a combined correlation matrix of the *ys* with *ys, xs* with *xs* and *ys* with *xs*.

In our current application, $q = 2$ and $p = 17$ and as $s = 2$ we have two pairs of canonical variates, i.e. cnv*f* 1, cnv*s*1 and cnv*f* 2, cnv*s*2. The standardised coefficients for the two pairs of variates and the corresponding canonical correlations calculated from the training data using all 17 predictors are presented in Table 3. The 17 variables (Table 4) are characterised by severe multicollinearity. Thus, the resulting model will be unstable with a reduced power of prediction. This is illustrated in Table 5 which shows the effect of changing the tolerance of the X-block variables on the subset selected for model specification, where:

$$\text{tolerance} = 1 - R_i^2, \tag{3}$$

Table 3. β estimates for the full 17 variable canonical correlation model.

	1st canonical variate	2nd canonical variate
Canonical correlations, r_c	0.92	0.79
log $K_i\alpha_1$	0.50	1.50
log $K_i\alpha_2$	0.57	−1.48
HFMOPAC	−0.35	−0.39
ShXY	−0.34	−0.068
MW	−8.38	5.95
AlogP98	0.42	−0.42
LogP	−0.024	−0.56
JDPSA1	0.076	−0.27
JPNSA3	−0.026	−0.97
JRPCS	−0.15	0.13
AC1	−0.24	−0.24
AC2	−2.21	2.73
AC9	−0.72	1.48
AC27	0.96	−1.59
ACl89	0.37	−0.37
SssCH2	4.98	−5.45
JX	0.12	1.55
Kappa3AM	−0.08	3.33
SC3C	5.75	−6.07

where R_i^2 is the coefficient of multiple determination of the ith x-variable with the remaining x-variables.

Specifying CCA models with high power of prediction

It was necessary to identify a set of canonical variates that have maximum power of prediction. This cannot be based on internal prediction since there are no N-fold cross-validation procedures available for CCA. It is possible, however, to obtain predictions for the two response variables (the $-\log K_i$ values for the α_1 and α_2 affinities) using a method akin to solving simultaneous equations [23]. These predictions can be used to identify robust QSAR models using the following procedure.

1. Using the training set data, CCA models were determined for each of the subsets of predictors identified in Table 4.
2. These models were then used to predict the α_1 and α_2 affinities of the test set compounds using the method of Ford and Salt [23].
3. The coefficients of determination (R^2) were calculated for the observed and predicted α_1 and α_2 affinities of the test set using the models specified in Table 4.
4. The estimates of R^2 for each model were ranked.
5. The model(s) with the greatest association (R^2) between the observed and predicted α_1 and α_2 affinities of the test set was (were) identified as that (those) with maximum power of prediction.

Table 4. Specified variables for different tolerance (TOL) values.

Variables		TOL							
		0.0001	0.001	0.01	0.1	0.2	0.28	0.3	0.4
1	HFMOPAC	X	X	X	X	X	X	X	X
2	SHXY	X	X	X	X				
3	MW	X							
4	ALOGP98	X	X	X					
5	LOGP	X	X	X	X	X	X	X	X
6	JDPSA1	X	X	X					
7	JPNSA3	X	X	X					
8	JRPCS	X	X	X	X	X	X		
9	AC1	X	X	X	X	X	X	X	X
10	AC2	X	X	X	X	X	X	X	
11	AC9	X	X	X	X	X	X	X	X
12	AC27	X	X	X	X	X	X		
13	ACL89	X	X	X	X	X	X	X	X
14	SSSCH2	X	X						
15	JX	X	X	X					
16	KAPPA3AM	X	X	X					
17	SC3C	X	X	X	X	X	X	X	X

Table 5. Pivoted variables and canonical correlation coefficients for the tolerance values of the X-block used in this study.

Tolerance	No. pivoted variables	R_C set	Training
0.0001	17	0.92	0.79
0.001	16	0.90	0.78
0.01	15	0.90	0.78
0.10	10	0.90	0.74
0.20	9	0.88	0.73
0.28	8	0.88	0.73
0.30	7	0.88	0.69
0.40	6	0.85	0.65

An optimal QSAR model was determined for both α_1 and α_2 affinity. The model predictions are plotted against the observed $-\log K_i$'s for both the test and training sets (Figure 4). They are in good agreement with the observations.

Predicting α_1 and α_2 affinities

The CCA model is defined below (Equations 4–7) in terms of standardised variables.

$$\text{cnvr}f1 = 0.62(-\log K_i\alpha_1) + 0.44(-\log K_i\alpha_2) \quad (4)$$

$$\text{cnvr}f2 = -1.45(-\log K_i\alpha_1) + 1.52(-\log K_i\alpha_2) \quad (5)$$

$$\begin{aligned}\text{cnv}s1 = {} & 0.21(\text{HFMOPAC}) - 0.05(\text{JRPCS}) \\ & + 0.31(\text{AC1}) + 0.56(\text{AC2}) + 0.11(\text{AC9}) \\ & - 0.28(\text{AC27}) + 0.001(\text{AC189}) \\ & + 0.16(\text{SC3C})\end{aligned} \quad (6)$$

$$\begin{aligned}\text{cnv}s2 = {} & -0.47(\text{HFMOPAC}) + 0.40(\text{JRPCS}) \\ & + 0.12(\text{AC1}) + 0.68(\text{AC2}) \\ & + 0.27(\text{AC9}) - 0.53(\text{AC27}) \\ & + 0.63(\text{AC189}) - 1.37(\text{SC3C})\end{aligned} \quad (7)$$

The loadings of the responses and predictors onto the variates (their correlations with each canonical variate) are presented in Table 6. Plots of the observed and predicted affinities are presented in Figure 4. The coefficients of determination for α_1, α_2 and pooled affinities are shown in Table 7. These results confirm that the 8-variable model constructed from the training set produces optimised predictions (maximum R^2_{pred}) of the test set data for all three prediction response sets (for α_1, α_2 and pooled), although prediction of α_1 affinity ($R^2_{\text{pred}} = 0.871$) is more reliable than that for α_2 ($R^2_{\text{pred}} = 0.657$).

Mapping the compounds in canonical variate space

The optimal CCA model was investigated further. The first canonical response variate (Equations 4 and 6) is a weighted sum of the α_1 and α_2 affinities, whereas the second (Equations 5 and 7) is a weighted difference. This result suggests that the CCA model can predict both affinity and selectivity. A plot of these two variates against each other confirms this proposition.

The response map (Figure 5) can be divided into nine mutually exclusive zones according to the combinations of values for the two affinities. The first canonical variate (cnv*r*1) is the weighted sum of both responses and therefore, for a compound to score high on cnv*r*1, it must have high affinity for both α_1 and α_2 (zones A, B and C). Conversely, if both activities are low then the compound will have a low score on this variate (zones G, H and I). If the compound has a high affinity for one of the responses and low for the other, or if they both have moderate values, then the compound will have a moderate score for cnv*r*1 (zones D, E and F). In contrast, cnv*r*2 is the weighted difference of

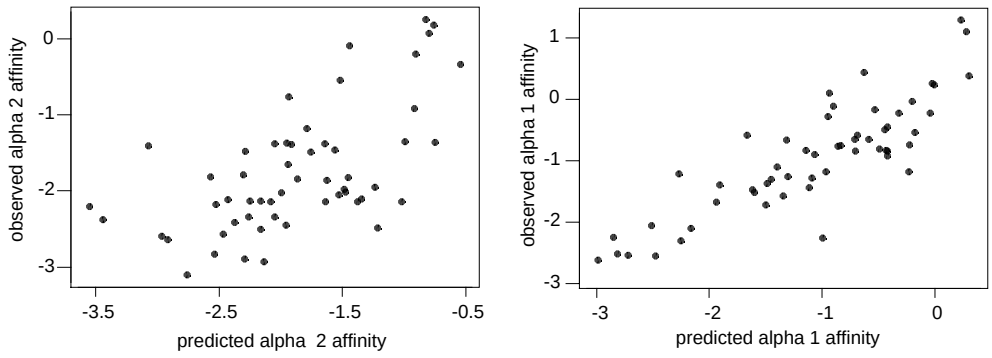


Figure 4. Plots of the predicted and observed affinities for the test set compounds using the 8-variable model (tolerance = 0.28).

Table 6. Loadings of the responses and predictors onto the canonical variates, cnv*r*1 and cnv*r*2.

Predictor	Loading onto cnv*r*1	Loading onto cnv*r*2
$-\log K_i\ \alpha_1$	**0.84**	−0.20
$-\log K_i\ \alpha_2$	**0.80**	0.29
HFMOPAC	−0.15	−0.58
JRPCS	−0.62	−0.17
AC1	**0.67**	0.09
AC2	**0.87**	0.14
AC9	**0.51**	0.10
AC27	−0.43	−0.18
AC189	**0.38**	0.12
SC3C	**0.81**	−0.36

Significant loadings ($P < 0.05$) emboldened.

the two responses and is of the general form $a(-\log act1) + b(-\log act2)$, where the constant a is negative and b is positive. Consequently, for a compound to score highly on cnv*r*2 it must have a high α_2 and a low α_1 affinity (zones C, F and I). To have a low score on this variate describes the opposite situation; low affinity for α_2 and high for α_1 (zones A, D and G). To have a moderate score on this variate, given that $a \approx b$ the two activities should be approximately the same (zones B, E and H). Thus, cnv*r*1 describes α_1 and α_2 affinity, whereas cnv*r*2 describes α_1 and α_2 selectivity.

To determine which particular zone a compound lies in we have to consider the combinations of activities described above. For example, zone A corresponds to high affinity for both α_1 and α_2 (cnv*r*1) and selectivity for α_1 but not for α_2. Clearly this condition cannot be satisfied and so no compounds appear in this zone of the response map, i.e., it is not feasible to find compounds in zone A. Zone F corresponds to either high affinity for one of the responses and low for the other or both moderate affinity (cnv*r*1) and high α_2 affinity and low α_1 affinity (cnv*r*2). High values for α_2 affinity and low values for α_1 satisfy these conditions and so zone F is feasible and corresponds to selectivity for α_2 activity but not for α_1. The full list of zones and their corresponding attributes is given below (Table 8).

Compounds with a combination of high affinity and selectivity for α_1 adrenoceptors will occupy zone D. Those with high affinity and selectivity for α_2 receptors will occupy zone F. Non-selective compounds with high affinity for both α_1 and α_2 receptors occupy zone B. Thus, the CCA model can predict those compounds with sufficient selectivity and affinity to be regarded as candidates for synthesis. It is therefore a useful model for virtual screening of these gene-targeted receptors.

Discussion and conclusions

The principal aim of this study, to predict the affinity and selectivity of arylpiperazinyl derivatives for α_2-ARs using robust QSAR models, has been achieved using a hypothesis generation strategy of selecting variables from a large set calculated using the Cerius 2 software and reducing this to a small set of relevant predictors, free from redundancy and unwanted multicollinearity. Two procedures, CR (Portsmouth fixed alpha) and CCA, have been used for this purpose. Both models comprise common descriptors (AC2, ACP, AC27, SC3C), but the CCA model requires

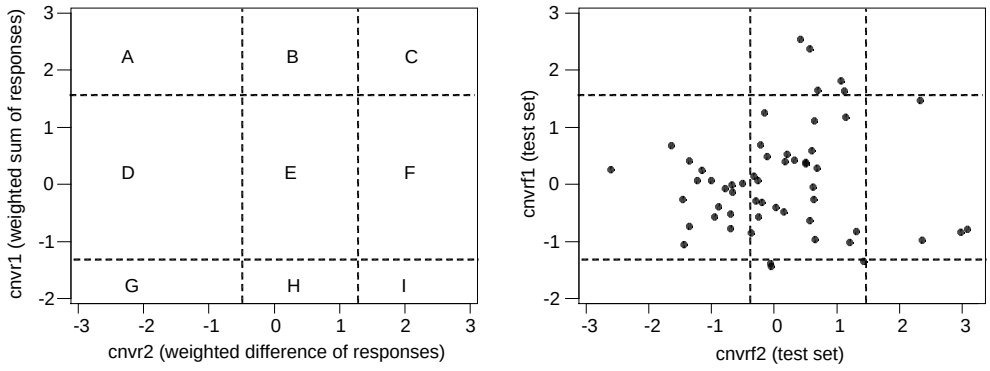


Figure 5. Prediction map for the test set compounds based on the first and second canonical variates (cnv*r*1 and cnv*r*2) calculated using the predictor set.

Table 7. Coefficients of determination for the α_1 and α_2 affinities observed and predicted for the test set.

Number of specified variables	X-block tolerance ($1-R^2$)	R^2 for α_1 affinity test set predictions	R^2 for α_2 affinity test set predictions	R^2 for pooled affinities
17	0.0001	0.803	0.461	0.696
16	0.001	0.795	0.557	0.752
15	0.01	0.699	−0.248	0.504
10	0.10	0.803	0.542	0.752
9	0.20	0.850	0.617	0.799
8	0.28	0.871	0.657	0.824
7	0.30	0.857	0.646	0.814
6	0.40	0.837	0.574	0.784

Table 8. Regions of the CCA map corresponding to different combinations of affinity and selectivity for α_1 and α_2-ARs.

Zone	Affinity combination
A	High α_1 and α_2 affinity, but selective for α_1 – not feasible
B	Both α_1 and α_2 affinity is high
C	High α_1 and α_2 affinity, but selective for α_2 – not feasible
D	High α_1 but low α_2 affinity, selective for α_1
E	Both α_1 and α_2 having moderate affinity
F	Low α_1 but high α_2 affinity, selective for α_2
G	Low α_1 and α_2 affinity, but selective for α_1 – not feasible
H	Low α_1 and α_2 affinity
I	Low α_1 and α_2 affinity, but selective for α_2 – not feasible

additional descriptors (HFMOPAC, JRPCS, AC1 and AC189) in order to specify affinity and selectivity. It is encouraging to observe that, with a single exception, all of the descriptors common to both the CR and CCA prediction models have the same sign in the affinity equations. The exception (SC3C) loads significantly onto both canonical variates (Table 6), and is therefore important for explaining both affinity and selectivity. This may explain its unstable behaviour in these two models.

Virtual screening for α_2 affinity

Internal prediction based on two-fold cross-validation suggested that predictions of affinity of use for virtual screening can be obtained using CR. The observed two-fold Q^2 had a mean value of 0.72 combined with a relatively tight distribution (small SD) and provided reasonable agreement ($r = 0.6$, $P < 0.001$) between test set prediction and observation. The mean value of Q^2 (0.72) gives a strong indication that not all of the properties necessary to explain affinity have been taken into account by the CR model and identification of these should lead to improved prediction. Even so, setting a threshold of –1.5 for $-\log K_i$ α_2 leads to a useful enrichment. Selection of compounds above this threshold would have reduced the number of candidates from 55 to 17. This reduced set would have included the seven most active molecules in the test set (Figure 3). Thus, the value of $-\log K_i$ $\alpha_2 = -1.5$ is a useful cut-off point for virtual screening based on the CR model. Finally, the QSAR model for α_2 affinity is constructed using only 2D descriptors, suggesting that affinity for α_2 ARs might be independent of the 3D conformational properties of this series of molecules.

Virtual screening for α_2 selectivity

The selectivity of a compound for α_1 or α_2-ARs can be judged using 2D maps based on the canonical variable space defined by cnv*r*1 and cnv*r*2. These maps are based on a CCA model that contains a number of the predictors used in the CR model. This consistency is encouraging and suggests that the models are generalised, capturing information about the population of arylpiperazinyl derivatives rather than sampling artifacts. At least some of the extra predictors are likely to be involved in determining the contrast between α_1 and α_2 activity, i.e. selectivity, a proposition supported for HFMOPAC by its significant loading onto cnv*r*2 (Table 6). The locations of predictions in classes D and F of the map (Table 8) indicate selectivity for α_1 and α_2, respectively and can

therefore be used for this purpose in virtual high throughput screening. In contrast to α_2 affinity, the α_2 selectivity model which deals with the contrasting affinities of α_1- and α_2-ARs is constructed using the 3D descriptors HF and JRPCS. This suggests that conformation may play a role in determining selectivity for these two receptor types.

Variable selection and generalised model specification

The principal task of model specification is to identify a model that is neither over- nor underfitted [4]. For robust QSAR models that are properly generalised but parsimonious, three key features are required. The predictor set should

1. contain only relevant variables each able to contribute additional information to the explanation of the response variable(s);
2. be free of redundancy, i.e., each predictor must contain some unique and relevant information;
3. have only an appropriate level of population multicollinearity that is useful in explaining the response, but be free of high levels of chance multicollinearities that have arisen from sampling.

This has been achieved in this study by adopting a set of procedures that addresses each of these requirements in turn. Relevant variables are easily identified from their associations with the Y-block. In this study, we have assessed this using Pearson's correlation coefficient, but more effective procedures that identify non-linear associations might have improved this aspect of the work. Redundancy was removed from the data matrix (whose initial number of variables greatly exceeded the rank of the matrix) using the UFS algorithm developed at Portsmouth and available at: www.cmd.port.ac.uk. However, it was first necessary to reduce the set to one containing only relevant predictors before undertaking UFS, in order to maintain computational speed and efficiency. The final problem of handling and optimising the multicollinearity of the predictor set is an altogether more complex and difficult task.

There are a number of sources of multicollinearity. Two of the most common and, for our purpose, important sources are those describing the population properties and those that arise as sampling artifacts. Population multicollinearities form an important feature of the properties of the molecules under investigation and contain important information that research scientists need to know about their set of molecules, particularly if the multicollinearities are associated with a feature related to the explanation of y. Sampling multicollinearities, however, are local phenomena of no interest other than their ability to lead to unstable regression models. They should always be identified [15] and removed from a QSAR model.

The instability problem arises during model specification because of the influence that multicollinearity has in inflating the errors of the β estimates. This occurs regardless of the source of the multicollinearity. However, if a population multicollinearity has some explanation of y, its β estimate based on the original, untransformed descriptors will also inflate and the t-statistics (calculated as the ratios of the β estimates and their SEs), which are used to assess the significance of the contributions of sets of multicollinear predictors, are likely to remain stable. Using these ideas, it may be possible to gain a better understanding of the problems that can arise when trying to remove multicollinearities in order to stabilise and generalise a QSAR model.

The most common approach to dealing with multicollinearities is to use a latent variable regression procedure such as CR [16, 22], PLS [24], PCR [25] or ridge regression [26, 27] in order to project the original variables comprising the predictor set onto eigenvectors, some of which reflect the population and some the sample multicollinearities. These are recognised because their weighted contributions from the original variables sum to a value close to zero (redundancy arises whenever this sum is equal to zero). The task is then to identify a small number of these short eigenvectors for which the eigenvalue, λ, tends to zero (typically $\lambda < 1$) that have the power to explain the Y-block and to include these, but only these, when specifying a robust QSAR model. Thus, the final set of X-block eigenvectors may include multicollinear predictors, but only if they lead to stable models in spite of the inflation of the SEs of the β estimates.

The processes of component (eigenvector) construction and selection are crucially important to the task of model specification. In this study, it has been assumed that component construction is

Table 9. Analysis of Variance (ANOVA) for the optimised CR regression with $\alpha = 0.3758$.

Source	Sum of squares	dof	Variance	*F* ratio	*P*
Model (C1,C5,C2)	38.0084	3	12.6695	44.37	< 0.00009
Error	13.9916	49	0.2855		
Total	52.0000	52			

$s = 0.53$, $R^2 = 0.73$, $R^2_{adj} = 0.72$, $Q^2_{LOO} = 0.74$.

simply a transformation of the original variables and has no implication for the degrees of freedom associated with a component. This is an unproven assumption. However, adoption of this argument leads to inclusion of the most useful components to result in a parsimonious and robust prediction model. This is supported by the fact that, once the construction criterion (α for CR) had been optimised (to give a value of 0.3758), the components selected on the basis of their contribution to the regression model determined using a conventional ANOVA approach identified the model with the highest internal consistency (LOO Q^2) and internal prediction (two-fold Q^2). Significantly, the second component (C5) to enter the model is a source of multicollinearity, since its length is short ($\rho = 0.0018$), reflecting the fact that its weighted linear combination of original variables tends to zero (Table 9) [23]. It has, however, a sizeable projection ($\tau = 8.81$) onto α_2 affinity (ρ and τ are respectively (i) the variance and (ii) the covariance with respect to the response variable, y, of a CR component). Moreover, this model resulted in useful external prediction when evaluated against the test set.

As a result of the model specification, the 6-variable space has been contracted to remove sampling artifacts, but retain all of the relevant population features (the signal) that explain the Y-block, including significant population multicollinearities such as those represented by C5 that are necessary for robust, reliable prediction. Finally, analysis of the residuals reveals no serious biases, suggesting that the model is well conditioned to the data and capable of robust prediction.

Acknowledgements

The authors wish to thank Professor Peter Goodford for initiating this study. Laura Maccari was supported by a postgraduate studentship funded by Molteni Farmaceutici. We are grateful for the helpful suggestions made by the referees during the vetting process.

References

1. Ford, A.P.D.W., Williams, T.J., Blue, D.R. and Clarke, D.E., Trends Pharmacol. Sci., 15 (1994) 167.
2. Hieble, J.P., Bylund, D.B., Clarke, A.E., Eikenberg, D.C., Langer, S.Z., Lefkowitz, R.J., Minneman, K.P. and Ruffolo, Jr. R.R., Pharmacol. Rev., 47 (1995) 266.
3. Hieble, J.P. and Ruffolo, Jr. R.R., Prog. Drug Res., 47 (1996) 81.
4. Lepor, H., J. Androl., 12 (1991) 389.
5. Caine, M., Urol. Clin. North Am., 17 (1990) 641.
6. Betti, L., Corelli, F., Floridi, M., Giannaccini, G., Maccari, L., Manetti, F., Strappaghetti, G. and Botta, M., J. Med. Chem., 46 (2003) 3555.
7. Clark, R.D., Michel, A.D. and Whiting, R.L., Prog. Med. Chem., 23 (1986) 1.
8. Berlan, M., Montastruc, J.L. and Lafontan, M., Trends Pharmacol. Sci., 13 (1992) 277.
9. Strader, C.D., Sigal, I.S. and Dixon, R.A., FASEB J., 3 (1989) 1825.
10. Strader, C.D., Sigal, I.S. and Dixon, R.A., Trends Pharmacol. Sci., 10 (Dec.Suppl.) (1989) 26.
11. Barbaro, R., Betti, L., Botta, M., Corelli, F., Giannaccini, G., Maccari, L., Manetti, F., Strappaghetti, G. and Corsano, S., J. Med. Chem., 44 (2001) 2118.
12. Betti, L., Botta, M., Corelli, F., Floridi, M., Giannaccini, G., Maccari, L., Manetti, F., Strappaghetti, G., Tafi, A. and Corsano, S., J. Med. Chem., 45 (2002) 3603.
13. Barbaro, R., Betti, L., Botta, M., Corelli, F., Giannaccini, G., Maccari. L., Manetti, F., Strappaghetti, G. and Corsano S., Bioorg. Med. Chem., 10 (2002) 361.
14. Catalyst, version 4.6: Accelrys, Inc., San Diego, CA.
15. Smellie, A., Teig, S.L. and Towbin, P., J. Comput. Chem., 16 (1995) 171.
16. Smellie, A., Kahn, S.D. and Teig, S.L., J. Chem. Inf. Comput. Sci., 35 (1995) 285.
17. Smellie, A., Kahn, S.D. and Teig, S.L., J. Chem. Inf. Comput. Sci., 35 (1995) 295.
18. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4 (1983) 187.
19. Whitley, D., Livingstone, D.J. and Ford, M.G., J. Chem. Inf. Comput. Sci., 40 (2000) 1160.
20. Belsley, D.A., Kuh, E. and Welsch, R.E., Regression Diagnostics, Identifying Influential Data and Sources of Collinearity, Wiley, New York, 1980, p. 104.
21. Malpass, J., Salt, D.W., Ford, M.G., Wynn, E.W. and Livingstone, D.J., In Van de Waterbeemd, H. (Ed.),

Methods & Principles in Medicinal Chemistry, VCH Publishers, Weinheim, Germany, 1995, pp. 163–189.
22. Stone, M. and Brooks, R.J., J. R. Statist. Soc. B, 52 (1990) 237; Wold, H., In Krishnaiah, P.R. (Ed.), Multivariate Analysis, Academic Press, New York, 1966.
23. Ford, M.G. and Salt, D.W., In van de Waterbeemd, H. (Ed.), Methods & Principles in Medicinal Chemistry, VCH Publishers, Weinheim, Germany, 1995, pp. 265–282.
24. Wold, H., In Krishnaiah, P.R., (Ed.), Multivariate Analysis, Academic Press, New York, 1966.
25. Livingstone D., Data Analysis for Chemists, Oxford University Press, Oxford, UK, 1995, p. 162.
26. Hoerl, A.E. and Kennard, R.W., Technometrics, 12 (1970) 55.
27. Hoerl, A.E., Kennard R.W. and Baldwin, K.F., Commun. Statist., 4 (1975) 105.