QSAR studies on BK channel activators

ARTICLE *in* BIOORGANIC & MEDICINAL CHEMISTRY · DECEMBER 2008

Impact Factor: 2.79 · DOI: 10.1016/j.bmc.2008.10.068 · Source: PubMed

Bioorganic & Medicinal Chemistry 17 (2009) 319–325

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

QSAR studies on BK channel activators

Alessio Coi [a], Francesca Lidia Fiamingo [a], Oreste Livi [a], Vincenzo Calderone [b], Alma Martelli [b], Ilaria Massarelli [c], Anna Maria Bianucci [a,*]

[a] Dipartimento di Scienze Farmaceutiche, Università di Pisa, Via Bonanno 6, 56126 Pisa, Italy
[b] Dipartimento di Psichiatria, Neurobiologia, Farmacologia e Biotecnologie, Università di Pisa, Via Bonanno 6, 56126 Pisa, Italy
[c] Dipartimento di Chimica e Chimica Industriale, Università di Pisa, Via Risorgimento 35, 56126 Pisa, Italy

ARTICLE INFO

Article history:
Received 24 June 2008
Revised 10 October 2008
Accepted 30 October 2008
Available online 5 November 2008

Keywords:
BK channel
QSAR
CODESSA

ABSTRACT

QSAR studies were developed on the basis of a dataset comprising BK channel activators previously synthesized and biologically assayed in our laboratory, in order to obtain highly accurate models enabling prediction of affinity toward the channel for New Chemical Entities (NCEs). Many molecular descriptors were computed by the CODESSA software. They were initially exploited in order to rationally split the available dataset into training and test set pairs, which supplied the basis for the development of QSAR models. Models were subjected to rigorous validation analysis based on the estimate of several statistical parameters, for the seek of the most accurate and simplest model enabling prediction of BK channel affinity.

1. Introduction

The large-conductance, calcium-activated potassium channels (BK, also termed BKCa, Slo, or MaxiK) are distributed in both excitable and non-excitable cells. They are involved in many cellular functions such as action potential repolarization, neuronal excitability, neurotransmitter release, hormone secretion, tuning of cochlear hair cells, innate immunity, and modulation of the tone of vascular, airway, uterine, gastrointestinal, and urinary bladder smooth muscle tissues.[1–3] BK channels characteristically respond to two distinct physiological stimuli, i.e. changes in membrane voltage and in cytosolic Ca^{2+} concentration.

The BK channels open in response to an increase in cytosolic Ca^{2+} concentration and membrane depolarization, resulting in an increase of K^+ efflux, which leads to rapid hyperpolarization of the excitatory membranes and reduces Ca^{2+} influx through voltage-dependent Ca^{2+} channels.

Then, the availability of exogenous compounds able to activate BK channels (usually named BK-activators) can guarantee an innovative pharmacological tool for the clinical management of many pathological states, due to a cell hyperexcitability, such as asthma, urge incontinence and bladder spasm, gastric hypermotility, neurological and psychiatric disorders.[1,2]

In earlier works by us several compounds were synthesized and subjected to in vitro pharmacological studies in order to evaluate their vasorelaxing effect related to the BK channel opening ability.[4–12] An investigation based on QSAR approaches was undertaken with the aim of rationalizing the BK-opening profile of 71 molecules that constitute the dataset on which predictive models were built starting from several different combinations of Training/Test (TR/TS) set pairs.

2. Results and discussion

2.1. Rational splitting of the dataset into training and test set pairs

Each molecule of the initial dataset was represented as a point in a multi-dimensional space defined by all the selected molecular descriptors. On this basis, the dataset was split into TR/TS set pairs, so that points representing both TR and TS sets were distributed within the whole descriptor space occupied by the entire dataset, and each point of the TS set was close to at least one point of the TR set. It ensures that the similarity principle is followed when activity is predicted on the TS set. Rational splitting was accomplished by using a Sphere-Exclusion type[13–17] algorithm, optimized 'in house' (data not published). The idea behind the classical sphere-exclusion algorithm is to select molecules, whose similarities with each of the other selected molecules are not higher than a defined threshold. Each selected molecule creates a hyper-sphere around itself, so that any candidate molecule inside the sphere is

Abbreviations: QSAR, quantitative and structure activity relationships; TR, training; TS, test; MLR, multiple linear regression; MM, Molecular Mechanics.

* Corresponding author. Tel.: +39 0502219544; fax: +39 0502219605.
E-mail address: bianucci@dcci.unipi.it (A.M. Bianucci).

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.10.068

excluded from the selection in the TR set and driven toward the TS set. The radius of the sphere (R) is an adjustable parameter, determining the number of compounds selected and the diversity among them. In particular $R = c(V/N)^{1/K}$ where, K is the number of descriptors (dimensionality of descriptor space), V is the total volume V occupied by the representative points in the normalized descriptor space, N is the number of molecules, and c is the dissimilarity level.[18] Different values in dissimilarity level enable identifying different TR/TS sets pairs. When the total volume V is set equal to one, the volume corresponding to one point is equal to 1/N.

Many molecular descriptors were computed by using CODESSA[19] and 444 of them were found to be shared by the 71 molecules of the dataset.

These descriptors were normalized according to the following formula:

$$X_{ij}^{n} = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}}$$

where X_{ij} and X_{ij}^{n} are the non-normalized and normalized *j-th* ($j = 1, ..., K$) descriptor values for compound i ($i = 1, ..., N$), correspondingly, and $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values for *j-th* descriptor. Thus, for descriptors, min $X_{ij}^{n} = 0$ and, max $X_{ij}^{n} = 1$.

Then Euclidean distances between pairs of molecules, among the 71 molecules comprise in the whole datasets, were calculated in a 444-dimensional space. The rational splitting was performed by using different values of *Similarity threshold* (S_{max}), ranging from 1.6 to 3.1, so that 20 TR/TS set pairs were generated. Their relative populations range from a minimum of 5 to a maximum of 24 molecules to be comprised in the TS set, which corresponds to about 1/3 of the dataset.

2.2. QSAR models and statistical analysis

The correlation identified by using CODESSA on each selected TR set, provided several equations based on different numbers of descriptors.[19]

It may be worth to recall here that MLR calculates QSAR equations by performing standard multivariable regression calculations, with multiple variables in a single equation. When correlations are searched by multiple linear regression (MLR), the optimal ratio between the number of descriptor exploited and molecules comprised in the TR set is suggested to be 1:5. Indeed, it is assumed that the variables belong to an orthogonal set, which is difficult to achieve in the reality; however a condition, ensuring that powerful predictive models are obtained, is given by a poor correlation between variables. In this perspective, the number of independent variables initially considered should be not higher than 1/5 of the number of known compounds in the TR set.[20] A higher ratio often leads to over-correlated equations, that in turn gives rise to poorly reliable predictions. What mentioned above defined the maximum number of descriptors allowed on the basis of the size of initial dataset.

As a first step, only models, for which the conditions $R^2 > 0.6$ and $q^2 > 0.5$ were satisfied on the TR set, were considered for further development, among all the ones based on the 20 TR/TS set pairs (Table 1).

Subsequently a further validation step was performed on each TS set. It allowed to discard several models from the ones that had satisfied the conditions on the TR set. Regressions were performed between experimental vs. predicted and predicted vs. experimental pIC_{50}s for molecules of the 20 corresponding TS sets.

Even though plotting both regressions might appear redundant, such plots often present different statistics. Some authors suggest that, when at least one of the two cases is satisfied, the model may be considered valid on the basis of the external TS set.[21] Statistical parameters were first calculated for the regression between experimental vs. predicted pIC_{50}s on the TS sets. Only one model, conventionally named **TR66** (as obtained from a TR set of 66 molecules) and based on 13 descriptors, passed this validation step. Its statistical parameters are reported in Table 2. The value of average absolute error ($\bar{E}$) was also calculated and resulted to be 0.1, also supporting the high accuracy of the model obtained.

Table 1
Statistical parameters (R^2 and q^2) found on the relevant TR sets for each QSAR model identified on the basis of the preliminary validation step. Maximum number of descriptors and similarity threshold, S_{max}, are also reported.

Model	n. Descriptors	S_{max}	R^2	q^2
TR66	13	1.60	0.73	0.56
TR65	13	1.65	0.73	0.57
TR64	12	1.68	0.73	0.59
TR63	12	1.70	0.74	0.59
TR62	12	1.85	0.74	0.59
TR61	12	1.85	0.78	0.62
TR60	12	2.00	0.78	0.58
TR59	11	2.05	0.76	0.60
TR58	11	2.25	0.82	0.69
TR57	11	2.25	0.75	0.63
TR56	11	2.27	0.75	0.60
TR55	11	2.30	0.78	0.63
TR54	10	2.33	0.77	0.62
TR53	10	2.35	0.78	0.63
TR52	10	2.50	0.75	0.55
TR51	10	2.55	0.78	0.65
TR50	10	2.62	0.75	0.61
TR49	9	2.65	0.74	0.58
TR48	9	2.80	0.78	0.62
TR47	9	3.10	0.81	0.61

Subsequently, statistical parameters of the regressions between predicted vs. experimental activities were computed for the **TR66** model, as a further validity check. The relevant values, found for the TS set, were: $R'^2 = 0.91$, $R_0'^2 = 0.91$, $k_0' = 1.01$, and $(R'^2 - R_0'^2)/R'^2 = 0.007$.

Furthermore, in order to apply the *parsimony principle*, attempts were made for decreasing the number of descriptors while retaining the level of performance. The *parsimony principle* implies that, among models characterized by comparable performances, the simplest one should be chosen for predictive purposes.[22]

As described before, the **TR66** model previously identified was obtained by exploiting the maximum number of molecular descriptors allowed for a TR set comprising 66 molecules. The search for 'reduced' models started from it and proceeded by decreasing the number of the descriptors, but they showed lower performance than the 13 descriptors model **TR66**, that is finally proposed as the best one (Equation in Table 3). Figure 1 shows the plot of predicted vs. experimental pIC_{50} values.

Moreover the **TR66** model was subjected to the response permutation test, also known as y-scrambling.[23,24] This procedure involves fitting several models (20 in the case treated here) on the same dependent variables (X block), but on a permutated response. If strong correlations are found between the descriptors characterizing the model and the randomized response, then the significance of the goodness of a selected QSAR model should be regarded as due to chance correlations. After the y-scrambling test, the **TR66** model was found to perform much better than any of their permuted models (Table 4). It means that the probability to obtain similar or better models using random numbers is null. This is a clear proof that the model is not affected by any chance correlation, and it is likely to depict true relationships. In particular, it can be observed a high difference between R^2 and q^2 values of the real model and the mean R^2 and q^2 of its permuted models being $|\Delta R^2| = 0.51$ and $|\Delta q^2| = 0.54$ (Table 4).

Table 2
Statistical parameters (R^2, R_0^2, $(R^2 - R_0^2)/R^2$, k_0) found for the **TR66** model, identified on the basis of TS validation.

					TR parameters		TS parameters			
Model	Similarity threshold	TS	TR	Descriptors	R^2	q^2	R^2	R_0^2	k_0	$(R^2 - R_0^2)/R^2$
TR66	1.6	5	66	13	0.73	0.56	0.91	0.91	0.99	0.0001

Table 3
QSAR equation referring to the 13 descriptors **TR66** model.

	X	ΔX	*t*-test	
0	−3.37E+01	1.54E+01	−2.194.3	Intercept
1	−2.16E−01	6.29E−02	−3.433.3	Max n–n repulsion for a C–N bond
2	5.88E−01	1.61E−01	3.645.9	Kier flexibility index
3	1.15E+01	5.00E+00	2.304.4	Min atomic orbital electronic population
4	6.87E+00	2.13E+00	3.221.6	ZX Shadow/ZX Rectangle
5	2.55E+00	8.27E−01	3.084.2	Avg bond order of a O atom
6	7.53E+01	1.49E+01	5.037.4	Max valency of a H atom
7	−8.17E+00	1.62E+00	−5.035.8	Max valency of a O atom
8	5.20E+02	2.04E+02	2.550.6	HACA-2/TMSA [Zefirov's PC]
9	3.14E+01	1.00E+01	3.137.0	Min partial charge for a C atom [Zefirov's PC]
10	−5.71E+00	2.33E+00	−2.448.5	YZ Shadow/YZ Rectangle
11	2.95E+01	1.01E+01	2.923.5	Max partial charge for a H atom [Zefirov's PC]
12	3.34E−02	1.64E−02	2.039.1	ZX Shadow
13	−4.88E−02	3.18E−02	−1.531.9	HACA-1 [Quantum-Chemical PC]

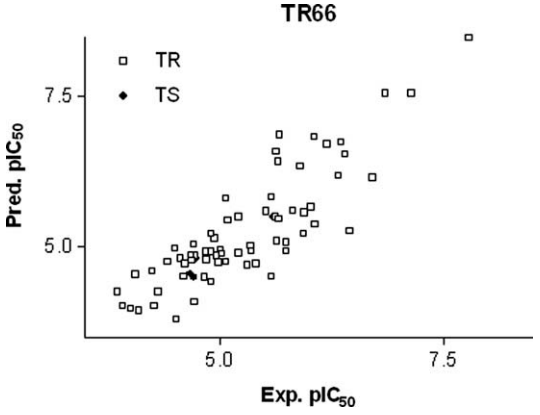


Figure 1. Plot of the predicted vs. experimental pIC_{50} values provided by **TR66** model on its TR and TS sets.

The 13-parameter model involves four of the six different class of molecular descriptors that may be computed within CODESSA. In particular, it relies on: one topological (Kier flexibility index), three geometrical (ZX Shadow/ZX Rectangle, YZ Shadow/YZ Rectangle, and ZX Shadow), three electrostatic (n. 8, 9, and 11), and six quantum-chemical (n. 1, 3, 5, 6, 7, and 13) descriptors (Table 3).

Most of the molecular descriptors can not be easily related to simple molecular features. However a general idea of their meaning may be captured according to what follows: (a) topological descriptors account for atomic connectivity in molecules; (b) geometrical descriptors require 3D-coordinates of the molecule and account for some properties such as molecular volume and surface areas. For example, shadow indices are geometrical indices related to the size and geometrical shape of the molecule; (c) electrostatic descriptors reflect the characteristics of the charge distribution using empirical partial charges calculated applying the approach proposed by Zefirov[19]; and (d) quantum-chemical, which are classified in different sub-types, refer to more complex, even more realistic, properties.

In particular, descriptors n. 3 and 13 can be classified as *Charge distribution-related* descriptors. These descriptors represent (or are directly related to) the quantum-chemically calculated charge distribution in the molecules, and therefore describe, in a very accurate way, the polar interactions between molecules or their chemical reactivity. Descriptors n. 5, 6, and 7 can be grouped as *Valency-related descriptors*. The descriptors of this class are related to the strength of intra-molecular bonding interactions and characterize the stability of the molecules, their conformational flexibility and other valency-related properties.[19] Finally, descriptor n. 1 can be defined as a *Quantum mechanical energy-related descriptor*. These descriptors characterize the total energy of the molecule in different energy scales and the intra-molecular energy distribution using different partitioning schemes.

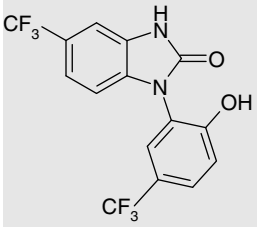
Table 4
Comparison between R^2 and q^2 (on TR set) found for the **TR66** model and its 20 permuted models (y-scrambling).

	R^2	q^2
y1	0.14	0.03
y2	0.27	0.04
y3	0.23	0.02
y4	0.33	0.03
y5	0.16	0.00
y6	0.34	0.05
y7	0.35	0.06
y8	0.29	0.03
y9	0.25	0.01
y10	0.13	0.04
y11	0.17	0.01
y12	0.20	0.00
y13	0.15	0.02
y14	0.18	0.01
y15	0.25	0.00
y16	0.22	0.00
y17	0.10	0.11
y18	0.15	0.01
y19	0.27	0.00
y20	0.17	0.00
Mean	0.22	0.02
Original **TR66**	0.73	0.56
Δ	0.51	0.54

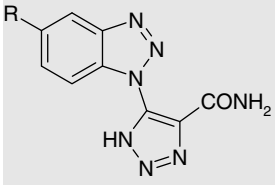
3. Conclusions

Understanding the relationships between the structural features of a series of compounds and their biological activity allows optimizing features responsible for the wanted interaction, when

Table 5
The 71 compounds, comprised in the dataset subjected to QSAR investigation.

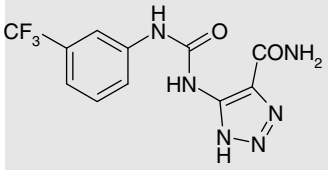


NS1619

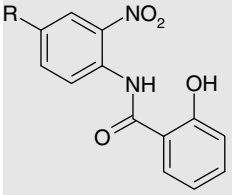


class "**II_2**":
a: R = H
b: R = CH_3
c: R = OCH_3
d: R = Cl
e: R = F
f: R = CF_3

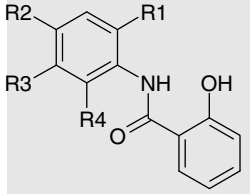
III_5b: R = C_6H_5



III_7

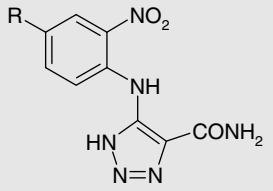


class "**IV_1**":
a: R = H
b: R = CH_3
c: R = Cl
d: R = F
e: R = CF_3
f: R = C_6H_5

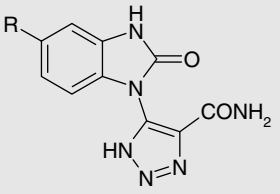


class "**V_12**":

a: R_1 = NO_2 R_2 = H R_3 = H R_4 = CH_3
c: R_1 = OCH_3 R_2 = NO_2 R_3 = H R_4 = H
d: R_1 = OH R_2 = NO_2 R_3 = H R_4 = H
e: R_1 = OCH_3 R_2 = H R_3 = NO_2 R_4 = H
f: R_1 = OH R_2 = H R_3 = NO_2 R_4 = H
g: R_1 = OH R_2 = H R_3 = Cl R_4 = H

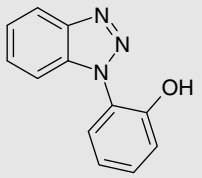


class "**I_2**":
a: R = H
b: R = CH_3
c: R = OCH_3
d: R = Cl

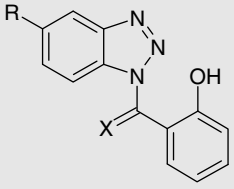


II_6b: R = CH_3

III_4b: R = C_6H_5

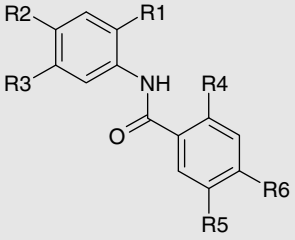


III_10



class "**IV_3**":
a: X = O; R = H
b: X = O; R = CH_3
c: X = O; R = Cl
d: X = O; R = F
e: X = O; R = CF_3
f: X = O; R = C_6H_5

IV_7: X = H; R = H



class "**V_16**":
a: R_1 = OH R_2 = H R_3 = CH_3
R_4 = OCH_3 R_5 = Cl R_6 = H
b: R_1 = OH R_2 = H R_3 = Cl
R_4 = OCH_3 R_5 = Cl R_6 = H

V_17: R_1 = OH R_2 = H R_3 = Cl
R_4 = OH R_5 = Cl R_6 = H

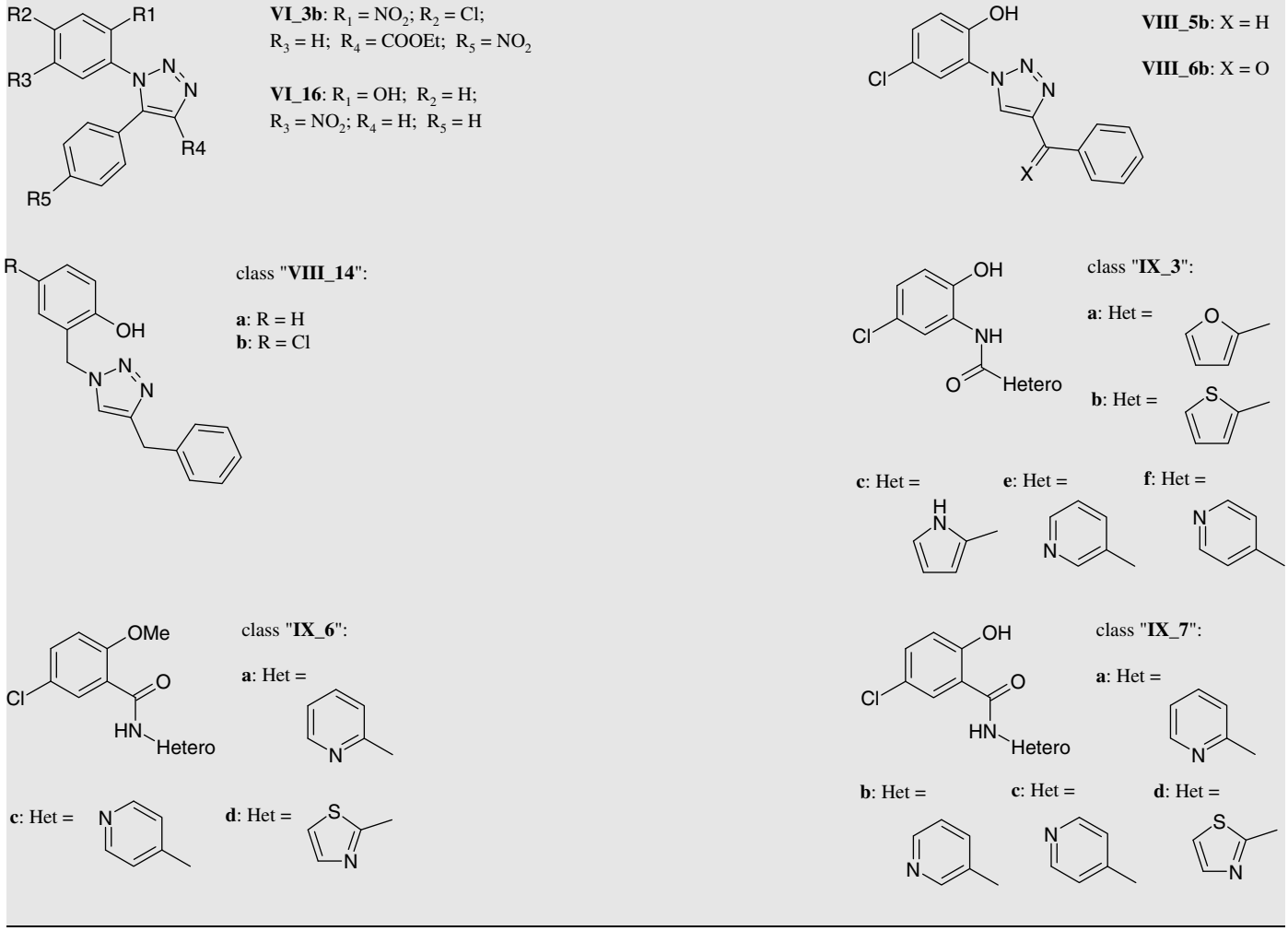
class "**X_4**":
a: R_1 = NO_2 R_2 = CH_3 R_3 = H R_4 = OH R_5 = Cl R_6 = H
b: R_1 = NO_2 R_2 = OH R_3 = H R_4 = OH R_5 = Cl R_6 = H
c: R_1 = OCH_3 R_2 = NO_2 R_3 = H R_4 = OH R_5 = Cl R_6 = H

class "**X_7**":
a: R_1 = OH R_2 = H R_3 = Cl R_4 = H R_5 = H R_6 = H
b: R_1 = OH R_2 = H R_3 = CH_3 R_4 = H R_5 = H R_6 = H
c: R_1 = OH R_2 = H R_3 = Cl R_4 = CH_3 R_5 = H R_6 = H
e: R_1 = OH R_2 = H R_3 = Cl R_4 = OCH_3 R_5 = H R_6 = H
f: R_1 = OH R_2 = H R_3 = CH_3 R_4 = OCH_3 R_5 = H R_6 = H
g: R_1 = OH R_2 = H R_3 = Cl R_4 = OEt R_5 = H R_6 = H
h: R_1 = OH R_2 = H R_3 = CH_3 R_4 = OEt R_5 = H R_6 = H
i: R_1 = OH R_2 = H R_3 = Cl R_4 = F R_5 = H R_6 = H

X_9a: R_1 = OH R_2 = H R_3 = Cl R_4 = H R_5 = Cl R_6 = H

class "**X_11**":
a: R_1 = OH R_2 = H R_3 = Cl R_4 = H R_5 = H R_6 = F
b: R_1 = OH R_2 = H R_3 = CH_3 R_4 = H R_5 = H R_6 = F

Table 5 (*continued*)



designing New Chemical Entities (NCEs). That allows removing undesirable molecules at early stages of their development, thus preventing waste of resources.

In this work, a QSAR investigation was performed on a dataset comprising 71 BK-activators, previously synthesized and assayed in house, with the aim of obtaining highly accurate models for the prediction of the affinity toward the channel. Among the number of models initially developed, a rigorous two-step validation analysis, based on commonly accepted statistical parameters and the response permutation test (y-scrambling), allowed identifying the simplest and most accurate model enabling to predict BK channel affinity. It can be regarded as a helpful tool for practical application in predicting BK channel activation properties.

As a future development the results presented here can be further improved by enlarging the initial dataset with other derivatives once they will in turn be designed, synthesized and biologically tested in our laboratory. This, consequently, will lead to an enlargement of the domain of validity of the model itself. Drug discovery, hence, could be an interactive dynamic process in which preliminary experimental results, analyzed by computational tools, could suggest how to optimize an initial class of derivatives. In turn, the experimental results collected from the newly designed derivatives act as a feedback for the computational studies and at the same time they give an enlargement of the initial content of information.

4. Methods

4.1. Database building and rational splitting into TR/TS set pairs

A dataset comprising 71 molecules, selected from earlier works by us and belonging to the classes of benzanilides, (benzo)triazoles, benzimidazolones, was subjected to the QSAR investigation presented here (Table 5). IC_{50} activation values were chosen as the 'target' property indicator to be exploited in QSAR analysis. For computational needs they were converted into the corresponding pIC_{50} [$pIC_{50} = Log(1/IC_{50})$] (Table 6). It may be worth to recall here that IC_{50} expresses the parameter of potency, representing the molar concentration of the tested compounds, which evokes half-reduction of the contractile tone induced by KCl 20 mM.[12] All the selected compounds had already been tested in our lab, thus ensuring the required homogeneity of biological data. Molecular structures of the entire dataset were optimized by means of a Molecular Mechanics (MM), according to a protocol previously described.[12] They were then subjected to semi-empirical Quantum-Mechanics calculations, by applying the AM1 hamiltonian, available in the MOPAC package[25], in order to obtain molecular properties that were then exploited for computing thermodynamic and quantum-chemical descriptors.

Molecular descriptors were calculated by the CODESSA program[19] and exploited for estimating structure similarities, expressed as Euclidean distances between pairs of molecules

Table 6
pIC_{50} values measured for the 71 compounds of the dataset.

Molecule	pIC_{50}	Molecule	pIC_{50}
V_16b[6]	8.49	X_11b[10]	4.95
V_16a[6]	7.56	IX_3b[9]	4.93
VIII_14a[8]	7.56	IX_7c[9]	4.93
V_12g[6]	6.87	X_9a[10]	4.93
IV_3b[5]	6.84	IX_6c[9]	4.91
VIII_5b[8]	6.75	X_7f[10]	4.90
V_12c[6]	6.72	IX_3c[9]	4.87
V_12a[6]	6.60	X_7a[10]	4.87
VI_3b[7]	6.55	II_2b[3]	4.86
V_17[6]	6.43	IX_3a[9]	4.82
V_12e[6]	6.35	IX_7b[9]	4.82
VIII6b[8]	6.19	IX_7d[9]	4.80
X_4c[10]	6.16	I_2f[2]	4.79
V_12d[6]	5.84	IX_6°[9]	4.77
IV_3a[5]	5.81	X_11a[10]	4.76
IV_3e[5]	5.67	III_10[4]	4.75
IV_3c[5]	5.61	I_2b[2]	4.73
IV_1b[5]	5.60	X_7e[10]	4.73
VI_16[7]	5.58	IV_7[5]	4.71
IV_3d[5]	5.52	I_2d[2]	4.61
IV_1a[5]	5.51	IX_3f[9]	4.57
X_7g[10]	5.51	I_2c[2]	4.55
IV_3f[5]	5.48	X_7c[10]	4.52
IV_1e[5]	5.46	X_7i[10]	4.52
X_4a[10]	5.39	IX_6d[9]	4.51
X_4b[10]	5.28	IX_7a[9]	4.51
IV_1f[5]	5.23	II_2f[3]	4.43
NS 1619[9]	5.23	II_2d[3]	4.26
IV_1c[5]	5.16	III_4b[4]	4.26
VIII14b[8]	5.11	III_7[4]	4.10
V_12f[6]	5.09	II_2e[3]	4.03
III_5b[4]	5.06	II_6b[3]	4.03
X_7b[10]	5.03	I_2a[2]	3.98
IX_3e[9]	4.99	II_2a[3]	3.95
IV_1d[5]	4.96	II_2c[3]	3.81
X_7h[10]	4.95		

among the 71 elements comprised in the dataset. On this basis, the dataset was split into several training/test (TR/TS) set pairs, by applying a sphere-exclusion type algorithm optimized in house. At first, the CODESSA output file was parsed and the exported descriptors were normalized. After that, some different similarity thresholds (S_{max}), ranging from 1.6 to 3.1, were applied, so generating 20 TR/TS set pairs within the dataset. Each pair was subsequently exploited for developing a number of models, each of them subjected to the required validation step.

4.2. CODESSA calculations

An heuristic procedure,[26] available in CODESSA, was applied to each one of the obtained TR sets, in order to accomplish a preliminary step enabling a first selection of molecular descriptors so that initial regression models were identified. This step has the aim of finding the most significant descriptors (from the standpoint of a single-parameter correlation), and which of them are highly inter-correlated. Such information enable optimizing the number of descriptors which have to be considered in the subsequent step of model development. First of all, descriptors are checked out to ensure that their variance, within the TR set is sufficient for structure–property correlation: descriptors showing low variance are discarded. Thereafter, all possible one-parameter regression models are checked out so that the poorly significant descriptors are removed. After that, the pair-correlation matrix of descriptors is calculated and the number of descriptors is further reduced by elimination of highly correlated descriptors. Two-parameter regression models with significant descriptors are subsequently obtained and ranked by the correlation coefficient R^2. A stepwise addition of further descriptors is performed as the final step, in order to find the best multi-parameter regression models with optimal values of the statistical parameters related to the adopted validation criteria (highest values of R^2, cross-validated R^2 (q^2), and Fisher criterion value F).

4.3. Statistical analysis and model validation

The predictive power of a QSAR model is reasonably verified by the comparison of experimental values of the 'target' property with theoretical values obtained by the model itself. The statistical parameter commonly exploited for comparison is the *correlation coefficient* (R^2). Most of the QSAR modeling methods implement the leave-one-out (or leave-many-out) cross-validation procedure (LOO-CV or LMO-CV), which consist of training a model over the whole dataset, but one (or some) element(s), and repeating the training step for all the possible combinations. The average R^2 (q^2) is so taken as a good validation parameter. Many authors had considered, up to recent times, high q^2 values as an indicator or, even, as the ultimate proof that the related QSAR model is highly predictive. However, several studies indicated that while high q^2 is a necessary condition for a model to have a high predictive power, it is not the sufficient one.[20,21]

On the other hand, the rigorous way for estimating the real predictive power of a model is to test it on a pool of compounds (TS set), provided it is fully disjoined from the TR set, being, this last one, the pool of compounds exploited for establishing structure–activity correlations. In addition molecules of TS set must belong to the so-called Applicability Domain (AD) of the QSAR model, defined by the TR set molecules. It means that the TR set must be able of properly sampling a chemical space where also TS set molecules have to be comprised. Therefore, in order to develop a QSAR model and validate it, the available dataset has to be rationally split into TR and TS sets. It was suggested that the TS set must include no less than five molecules, whose activities and structures must cover the range of activities and structures of compounds from the TR set. This requirement is necessary for obtaining reliable statistics for comparison between the experimental and activities predicted by the model for such compounds.[21] Thus, in addition to a high q^2, also high R^2 between the predicted and experimental activities of the TS molecules has to be included in the validation criteria. Golbraikh and Tropsha introduced the concept of 'ideal' QSAR model and formulated a set of criteria for assessing the reliability of a 'real' QSAR model on the basis of its closeness to the 'ideal' one.[21] Based on what has been highlighted above, the validation step has to be accomplished for each TR/TS set pairs, taking into account the conditions reported below, which have to be simultaneously satisfied:

criteria on the TR set: $R^2 > 0.6$, $q^2 > 0.5$,
criteria on the TS set: $R^2 > 0.6$, $0.85 < k_0 < 1.15$, $(R^2 - R_0^2)/R^2 < 0.1$

where R^2 is the correlation coefficient of the regression line between the predicted vs. experimental pIC_{50}, R_0^2 is the correlation coefficient of the same regression line forced through the origin and k_0 is the slope of this line.

References and notes

1. Calderone, V. *Curr. Med. Chem.* **2002**, *9*, 1385.
2. Wu, S.-N. *Curr. Med. Chem.* **2003**, *10*, 649.
3. Ghatta, S.; Nimmagadda, D.; Xu, X.; O'Rourke, S. *Pharmacol. Ther.* **2006**, *110*, 103.
4. Biagi, G.; Calderone, V.; Giorgi, I.; Livi, O.; Scartoni, V.; Baragatti, B.; Martinotti, E. *Eur. J. Med. Chem.* **2000**, *35*, 715.
5. Baragatti, B.; Biagi, G.; Calderone, V.; Giorgi, I.; Livi, O.; Martinotti, E.; Scartoni, V. *Eur. J. Med. Chem.* **2000**, *35*, 949.

6. Biagi, G.; Calderone, V.; Giorgi, I.; Livi, O.; Scartoni, V.; Baragatti, B.; Martinotti, E. *Il Farmaco* **2001**, *56*, 841.
7. Biagi, G.; Giorgi, I.; Livi, O.; Scartoni, V.; Barili, P. L.; Calderone, V.; Martinotti, E. *Il Farmaco* **2001**, *56*, 827.
8. Biagi, G.; Giorgi, I.; Livi, O.; Nardi, A.; Calderone, V.; Martelli, A.; Martinotti, E.; Salerni, O. L. *Eur. J. Med. Chem.* **2004**, *39*, 491.
9. Biagi, G.; Calderone, V.; Giorgi, I.; Livi, O.; Martinotti, E.; Martelli, A.; Nardi, A. *Il Farmaco* **2004**, *59*, 397.
10. Calderone, V.; Giorgi, I.; Livi, O.; Martinotti, E.; Mantuano, E.; Martelli, A.; Nardi, A. *Eur. J. Med. Chem.* **2005**, *40*, 521.
11. Calderone, V.; Fiamingo, F. L.; Giorgi, I.; Leonardi, M.; Livi, O.; Martelli, A.; Martinotti, E. *Eur. J. Med. Chem.* **2006**, *41*(6), 761.
12. Calderone, V.; Coi, A.; Fiamingo, F. L.; Giorgi, I.; Leonardi, M.; Livi, O.; Martelli, A.; Martinotti, E. *Eur. J. Med. Chem.* **2006**, *41*(12), 1421.
13. Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J. *Quant. Struct.-Act. Relat.* **1996**, *15*, 285.
14. Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. *J. Mol. Graphics Modell.* **1997**, *15*, 372.
15. Nilakantan, R.; Bauman, N.; Haraki, K. S. *J. Comput. Aided Mol. Design* **1997**, *11*, 447.
16. Clark, R. D. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181.
17. Gobbi, A.; Lee, M. L. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 317.
18. Golbraikh, A. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 414.
19. Katritzky, A.R.; Lobanov, V.S.; Karelson, M. CODESSA: Reference Manual; Version 2; University of Florida, 1994.
20. Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238.
21. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269.
22. Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1.
23. Van der Voet, H. *Chemom. Intell. Lab. Syst* **1994**, *25*, 313.
24. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and MegaVariate Data Analysis — Principles and Applications*; Umetrics AB: Umea, Sweden, 2001.
25. MOPAC, Version 7.0, Stewart, J.J.P. Fujitsu Limited, Tokyo, Japan.
26. Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley: New York, 2000.