

Computational neural network analysis of the affinity of *N*-*n*-alkylnicotinium salts for the $\alpha 4\beta 2^*$ nicotinic acetylcholine receptor

FANG ZHENG, GUANGRONG ZHENG, A. GABRIELA DEACIUC, CHANG-GUO ZHAN, LINDA P. DWOSKIN, & PETER A. CROOKS

Department of Pharmaceutical Sciences, College of Pharmacy, University of Kentucky, Lexington, Kentucky 40536, USA

(Received 31 August 2007; accepted 14 December 2007)

Abstract

Based on an 85 molecule database, linear regression with different size datasets and an artificial neural network approach have been used to build mathematical relationships to fit experimentally obtained affinity values (K_i) of a series of *mono*- and *bis*-quaternary ammonium salts from [^3H]nicotine binding assays using rat striatal membrane preparations. The fitted results were then used to analyze the pattern among the experimental K_i values of a set of *N*-*n*-alkylnicotinium analogs with increasing *n*-alkyl chain length from 1 to 20 carbons. The affinity of these *N*-*n*-alkylnicotinium compounds was shown to parabolically vary with increasing numbers of carbon atoms in the *n*-alkyl chain, with a local minimum for the C_4 (*n*-butyl) analogue. A decrease in K_i value between C_{12} and C_{13} was also observed. The statistical results for the best neural network fit of the 85 experimental K_i values are $r^2 = 0.84$, $\text{rmsd} = 0.39$; $r_{\text{cv}}^2 = 0.68$, and $\text{loormsd} = 0.56$. The generated neural network model with the 85 molecule training set may also be of value for future predictions of K_i values for new virtual compounds, which can then be identified, subsequently synthesized, and tested experimentally.

Keywords: *N*-*n*-alkylnicotinium salts, nicotinic acetylcholine receptor, binding affinity, neural network, linear regression

Introduction

Nicotine, the major alkaloid in tobacco, is the addictive compound that maintains tobacco smoking behavior [1–2]. Tobacco smoking is the number one cause of preventable mortality, and is responsible for over 4 million smoking-related deaths world-wide each year [3]. Although 70% of the 50 million people that smoke tobacco have attempted to quit, only 3% of those individuals maintain cessation for a period of one year using currently available therapies. Consequently, relapse rates for tobacco smoking continue to be high, indicating that novel smoking cessation therapies are needed [3].

Nicotine produces its effects on the central nervous system (CNS) by interacting with nicotinic acetylcholine receptors (nAChRs) that are essential for synaptic transmission. nAChRs consist of five protein

subunits which transverse the neuronal cell membrane. The most common nAChR subtype is $\alpha 4\beta 2^*$, which accounts for over 90% of the high-affinity [^3H]nicotine binding sites in brain [4]. $\alpha 4\beta 2^*$ nAChRs have been recognized as a major therapeutic target for mediating several CNS pathologies and diseases, including tobacco dependence [3,4].

A variety of $\alpha 4\beta 2^*$ agonists and antagonists have been discovered [5,6]. Previous research [7–13] in our laboratories has led to the discovery of a new class of nAChR antagonists composed of a series of *mono*- and *bis*-quaternary ammonium salts [13]. Among these are the *N*-*n*-alkylnicotinium salts that vary in the length of their *N*-*n*-alkyl chains [11]. After attachment of an *n*-alkyl chain to the pyridine nitrogen atom of the nicotine molecule, the properties of the resulting novel compounds are significantly altered. The pK_a value

Correspondence: P. A. Crooks, College of Pharmacy, University of Kentucky, Lexington, KY 40536-0086, USA. Tel: 1 859 257 1718. Fax: 1 859 257 7585. E-mail: pcrooks@email.uky.edu

of the nitrogen of the pyrrolidine ring drops from 8.5 to about 6 [14], and the compound is converted from an agonist to an antagonist at nAChRs [11,12]. These novel compounds exhibit potent and competitive inhibition of nAChR subtypes mediating S-(-)-nicotine-evoked dopamine release from dopaminergic nerve terminals in superfused rat striatal slices [11], and may have potential as smoking cessation agents, due to their selective antagonist activity at these nAChR subtypes [11].

An interesting phenomenon observed in the structure-activity profile of the *N*-n-alkylnicotinium series of compounds is that the experimentally measured K_i values obtained from the [^3H]nicotine binding assay (which probes the $\alpha 4\beta 2^*$ nAChR subtype) afforded a greater diversity in pharmacological response than might be expected for a homologous series of compounds where the *N*-n-alkyl chain length varies from C_1 to C_{20} . However, these data can be mined to identify associations that can then provide insights for future drug design studies. Generally speaking, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Technically, data mining is the process of discovering previously unknown relationships (correlations) or hidden patterns among a group of data in a database [15]. In this study, linear regression and artificial neural network approaches were used to analyze the pattern among the experimentally determined binding affinities (K_i values) at the $\alpha 4\beta 2^*$ nAChR subtype for a set of *N*-n-alkylnicotinium salts that varied in the length of their *N*-n-alkyl chains. The generated quantitative structure-activity relationship (QSAR) models are also valuable for predicting the affinity of hypothetical molecules as antagonists at $\alpha 4\beta 2^*$ nAChR subtypes which can then be identified as structures of interest, and therefore seriously considered for synthesis and pharmacological evaluation.

Materials and methods

Experimental database

A database of eighty-five *mono*- and *bis*-quaternary ammonium compounds with experimentally determined K_i values from the [^3H]nicotine binding assay was available. The [^3H]nicotine binding assay utilized rat striatal membrane preparations to probe the interaction of the analogs with the $\alpha 4\beta 2^*$ nAChR subtype [11]. Of the 85 molecules in the database, the K_i values of 7 of the molecules were $\leq 0.1 \mu\text{M}$, 13 molecules had K_i values in the range $0.1-1 \mu\text{M}$, 26 molecules had K_i values in the range $1-10 \mu\text{M}$, and 39 molecules had K_i values $\geq 10 \mu\text{M}$. This database was utilized for the linear and non-linear regression analyses (Table I).

Generation of molecular descriptors

Molecular modeling was carried out with the aid of the Sybyl discovery software package [16a]. This software was used to construct the initial molecular structures utilized in the geometry optimization (energy minimization) for all molecules involved in this study. The geometry optimization was first performed using a molecular mechanics (MM) method with the Tripos force field and the default convergence criterion, which was then followed by a semi-empirical molecular orbital (MO) energy calculation at the PM3 level [16b].

A set of 530 descriptors, including OD, 1D, 2D and 3D whim descriptors, were calculated by the DRAGON program [17] for these optimized molecules. In addition, a variety of other molecular properties, such as polar volume and polar surface area, were produced by use of the Tripos software package [16a]. Other steric descriptors were measured from the optimized 3D molecular structures, including the intra-molecular distance between the pyridine ring nitrogen and the pyrrolidine ring nitrogen in the nicotine derived molecules. The PM3 method was used to determine the LUMO, HOMO energies, dipole moments, and atomic charges of each molecule [16]. Pre-filtering for constant and pair-wise correlation (>0.95) descriptors were performed, and a stepwise procedure was carried out to select variables from the remaining 143 descriptors.

Target properties

Experimentally determined K_i values of the synthesized quaternary ammonium analogs were measured according to the procedure described by Dwoskin et al. [11]. The $\log(1/K_i)$ or pK_i (with K_i value in μM) was used as the target property for performing linear regression and neural network analyses.

Linear regression analysis

Linear least squares regression and multiple linear regression (MLR) analyses were performed utilizing an in-house Fortran77 program. Starting from the entire set of descriptors, variable selection by a forward and reverse stepwise regression procedure was performed, in which forward selection was followed by backward elimination of variables, resulting in an equation in which only variables that significantly increased the predictability of the dependent variable were included [18a].

Artificial neural network (ANN) analysis

Feed-forward, back-propagation-of-error networks were developed using a neural network C program

Table I. Structures, experimentally determined pK_i values (with K_i value in μM) from [^3H]nicotine binding assays, and pK_i values (with K_i value in μM) calculated by the NN731 model and its leave-one-out validation results for 85 quaternary ammonium salts[†].

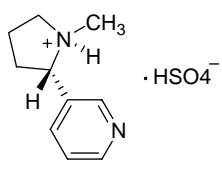
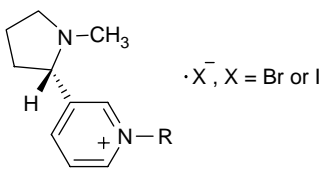
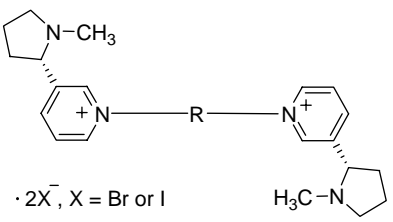
No.	Cmpd Name	R	pK_i (Expt)	pK_i (NN)	pK_i (NNLOO)
Nicotine					
					
1	NIC		2.7	2.59	2.44
<i>N</i> -Alkylnicotinium Salts					
					
2*	NMNI	CH ₃	-0.55	-0.36	-0.26
3*	NENI	n-C ₂ H ₅	-0.02	-0.37	-0.65
4*	NPNI	n-C ₃ H ₇	-1.33	-0.59	-0.25
5*	NnBNI	n-C ₄ H ₉	-1.04	-0.60	-0.51
6*	NHxNI	n-C ₆ H ₁₃	0.28	-0.61	-0.84
7*	NHpNI	n-C ₇ H ₁₅	-0.32	-0.59	-0.67
8*	NONI	n-C ₈ H ₁₇	-1.29	-0.44	-0.24
9*	NNNI	n-C ₉ H ₁₉	0.08	-0.25	-0.32
10*	NDNI	n-C ₁₀ H ₂₁	1.05	0.23	0.01
11*	NDDNI	n-C ₁₂ H ₂₅	0.85	0.70	0.56
12*	GZ511A	n-C ₁₃ H ₂₇	-0.42	-0.29	-0.38
13*	GZ511B	n-C ₁₄ H ₂₉	-0.85	-0.35	0.01
14*	GZ512A	n-C ₁₅ H ₃₁	0.96	0.65	0.54
15*	GZ512B	n-C ₁₆ H ₃₃	1.07	0.81	0.74
16*	GZ521A	n-C ₁₇ H ₃₅	1.28	1.40	1.41
17*	GZ521B	n-C ₁₈ H ₃₇	1.85	1.46	1.25
18*	GZ522B	n-C ₂₀ H ₄₁	0.89	1.22	2.14
19	NONB-7M	(CH ₂) ₆ CH(CH ₃) ₂	-0.37	-0.21	-0.15
20	NCyNB-4	cyclobutyl-(CH ₂) ₆ CH(CH ₂) ₃	-0.66	-0.49	-0.47
21	NCyNB-5	cyclopentyl-(CH ₂) ₆ CH(CH ₂) ₄	-0.74	-0.47	-0.84
22	NCyNB-6	cyclohexyl-(CH ₂) ₆ CH(CH ₂) ₅	-0.86	-0.41	0.13
23	NBzNB	CH ₂ C ₆ H ₅	-0.43	-0.33	-0.22
24	NANI	CH ₂ CH=CH ₂	-0.32	-0.34	-0.33
25	NONB-3c	cis-(CH ₂) ₂ CH=CH(CH ₂) ₃ CH ₃	1.10	-0.56	-0.77
26	NONB-3t	trans-(CH ₂) ₂ CH=CH(CH ₂) ₃ CH ₃	-0.65	-0.50	-0.47
27	NONB-7e	(CH ₂) ₆ CH=CH ₂	0.34	0.18	0.24
28	NONB-3y	(CH ₂) ₂ CC(CH ₂) ₃ CH ₃	0.70	0.14	-0.39
29	NONB-6e7m	(CH ₂) ₅ CH=CH(CH ₃) ₂	0.35	-0.39	-0.49
30	NDNB-4c	cis-(CH ₂) ₃ CH=CH(CH ₂) ₄ CH ₃	-0.87	0.05	0.29
31	NDNB-4t	trans-(CH ₂) ₃ CH=CH(CH ₂) ₄ CH ₃	0.49	0.69	0.49
32	NDNB-9e	(CH ₂) ₈ CH=CH ₂	1.30	1.51	0.41
33	NDNB-3y	(CH ₂) ₂ CC(CH ₂) ₅ CH ₃	0.24	0.70	0.64
34	NUNB-10e	(CH ₂) ₉ CH=CH ₂	0.80	0.77	1.17
<i>bis-N,N'</i> -Alkylnicotinium Salts					
					
35	bNHxI	(CH ₂) ₆	-1.30	-1.09	-0.87
36	bNOI	(CH ₂) ₈	-0.19	-0.11	-0.07
37	bNNB	(CH ₂) ₉	-0.70	-0.63	-0.24

Table I – continued

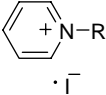
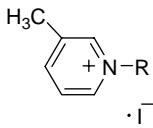
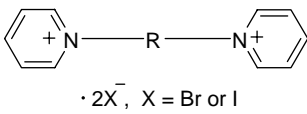
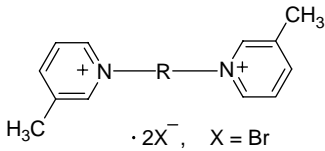
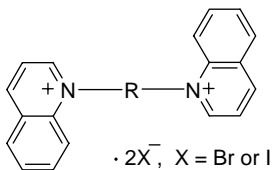
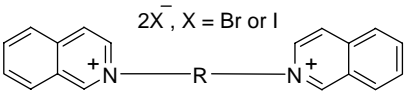
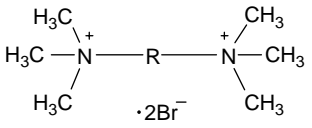
No.	Cmpd Name	R	p <i>K</i> _i (Expt)	p <i>K</i> _i (NN)	p <i>K</i> _i (NNLOO)
38	bNDI	(CH ₂) ₁₀	0.48	-0.01	-0.13
39	bNUB	(CH ₂) ₁₁	0.37	0.05	-0.19
40	bNDDB	(CH ₂) ₁₂	-0.29	0.00	0.63
<i>N</i> -Alkylpyridinium Salts					
					
41*	NMPI	CH ₃	-1.30	-1.27	-1.25
42*	NEPI	n-C ₂ H ₅	-1.46	-1.21	-1.17
43*	NPrPI	n-C ₃ H ₇	-1.63	-1.38	-1.37
44*	NBuPI	n-C ₄ H ₉	-0.98	-1.13	-1.14
45*	NPePI	n-C ₅ H ₁₁	-0.94	-1.30	-1.36
46*	NHxPI	n-C ₆ H ₁₃	-0.96	-1.17	-1.16
47*	NHpPI	n-C ₇ H ₁₅	-1.18	-1.33	-1.34
48*	NOPI	n-C ₈ H ₁₇	-1.30	-1.20	-1.15
49*	NNPI	n-C ₉ H ₁₉	-1.28	-1.39	-1.38
50*	NDPI	n-C ₁₀ H ₂₁	-1.22	-1.47	-1.48
51*	NUPI	n-C ₁₁ H ₂₃	-1.23	-1.51	-1.53
52*	NDDPI	n-C ₁₂ H ₂₅	-2.08	-1.53	-1.51
53*	NPeDPI	n-C ₁₅ H ₃₁	-1.58	-1.28	-1.17
54	NAPB	CH ₂ CH=CH ₂	-1.75	-1.76	-1.26
55	JTO	<i>cis</i> -CH ₂ CH=CH(CH ₂) ₄ CH ₃	-1.12	-0.97	-0.85
56	JCO	<i>trans</i> -CH ₂ CH=CH(CH ₂) ₄ CH ₃	-1.28	-1.25	-1.24
57	LO	CH ₂ CC(CH ₂) ₄ CH ₃	-1.24	-1.44	-1.45
58	MO	(CH ₂) ₆ CH=CH ₂	-1.05	-1.33	-1.35
59	LN	CH ₂ CC(CH ₂) ₅ CH ₃	-1.46	-1.32	-1.28
60	JTN	<i>cis</i> -CH ₂ CH=CH(CH ₂) ₅ CH ₃	-1.4	-1.14	-1.00
61	JCN	<i>trans</i> -CH ₂ CH=CH(CH ₂) ₅ CH ₃	-1.38	-1.36	-1.37
62	LD	CH ₂ CC(CH ₂) ₆ CH ₃	-0.98	-1.08	-1.00
63	MU	(CH ₂) ₉ CH=CH ₂	-1.09	-0.93	-0.80
64	NU	(CH ₂) ₉ CCH	-0.93	-1.06	-0.48
<i>N</i> -Alkyl-3-picolinium Salts					
					
65*	NNPiI	n-C ₉ H ₁₉	-1.79	-1.53	-1.50
66*	NDPiI	n-C ₁₀ H ₂₁	-1.41	-1.48	-1.51
<i>bis-N,N'</i> -Alkylpyridinium Salts					
					
67	bPHxI	(CH ₂) ₆	-1.99	-1.72	-1.63
68	bPOI	(CH ₂) ₈	-1.54	-1.66	-1.69
69	bPNB	(CH ₂) ₉	-1.38	-1.41	-1.33
70	bPDI	(CH ₂) ₁₀	-1.27	-1.55	-1.59
71	bPUB	(CH ₂) ₁₁	-1.15	-1.69	-1.73
72	bPDDB	(CH ₂) ₁₂	-0.96	-1.65	-1.70
<i>bis-N,N'</i> -Alkyl-3-picolinium Salts					
					
73	bPiNB	(CH ₂) ₉	-1.90	-1.68	-1.55
74	bPiUB	(CH ₂) ₁₁	-1.84	-1.44	-0.71

Table I – continued

No.	Cmpd Name	R	pK_i (Expt)	pK_i (NN)	pK_i (NNLOO)
75	bPiDDB	$(CH_2)_{12}$	-1.69	-1.65	-1.69
<i>bis-N,N'</i> -Alkylquinolinium Salts <div style="text-align: center;">  </div>					
76	bQHxI	$(CH_2)_6$	-1.59	-1.35	-1.08
77	bQNB	$(CH_2)_9$	-1.35	-1.73	-1.87
78	bQUB	$(CH_2)_{11}$	-1.58	-1.70	-1.75
<i>bis-N,N'</i> -Alkylisoquinolinium <div style="text-align: center;">  </div>					
79	bIQHxI	$(CH_2)_6$	-1.25	-1.27	-1.24
80	bIQOI	$(CH_2)_8$	-1.82	-1.45	-1.27
81	bIQNB	$(CH_2)_9$	-0.90	-0.93	-1.03
82	bIQUB	$(CH_2)_{11}$	-0.96	-0.52	0.04
83	biQDDB	$(CH_2)_{12}$	-0.79	-1.24	-1.47
<i>bis-Quaternary Ammonium Salts</i> <div style="text-align: center;">  </div>					
84	HEX	$(CH_2)_6$	-1.34	-1.40	-1.45
85	DEC	$(CH_2)_{12}$	-0.76	-1.21	-1.45

* Compounds were used in the 32 molecule dataset for multilinear regression analysis; † All compounds and experimental data in this table were generated in the laboratories of Dr. Peter A. Crooks and Dr. Linda P. Dvoskin. Most of these data have been published in References 7–13 and 20. Compounds 12–18 and their K_i values have not been published previously.

[18]. Network weights ($W_{ji}(s)$) for a neuron “j” receiving output from neuron “i” in the layer “s” were initially assigned random values between -0.5 and $+0.5$. The sigmoidal function was chosen as the transfer function that generates the output of a neuron from the weighted sum of inputs from the preceding layer of units. Consecutive layers were fully interconnected; there were no connections within a layer or between the input and the output. A bias unit with a constant activation of unity was connected to each unit in the hidden and output layers.

The input vector was constructed from the set of descriptors for each molecule in the database, as generated by the previous steps. All descriptors and targets were normalized to the $[0,1]$ interval. The network was configured with one or more hidden layers. To find the optimal number of neurons in the hidden layer, neural network architectures with different inputs and hidden neurons, respectively, were trained. During the ANN learning process, each

compound in the training set was iteratively presented to the network, i.e. the input vector of the chosen descriptors in normalized form for each compound was fed to the input units, and the network’s output was compared with the experimentally determined “target” value. During one “epoch”, all compounds in the training set were presented, and weights in the network were then adjusted on the basis of the discrepancy between network outputs and observed pK_i values by back-propagation, using the generalized delta rule [18].

Quality evaluation of regression analyses

A key point in any data mining process is quality evaluation of a mathematical analysis. It serves two purposes: i.e. to help identify the mathematical model that best represents the pattern in training data, and to predict how well the final model will work in the future [15]. As the total number of data points used in the

current study was less than 100, to obtain a precise estimate, the models generated with linear or non-linear mathematical approaches were validated by leave-one-out (LOO or loo) cross validation. The models were assessed by the Pearson correlation coefficient r^2 , root mean square deviation (rmsd), and predictive r_{cv}^2 , which is defined as:

$$r_{cv}^2 = \frac{SD - PRESS}{SD} \quad (1)$$

where SD is the sum of squared deviations of each measured pK_i value from its mean, and $PRESS$ is the predictive sum of squared differences (the sum of squared differences between actual and predicted values). The final mathematical relationship was determined as the one having the smallest rmsd error of LOO validation for various linear or non-linear regression analyses.

Results and discussion

Results from linear regression analysis

The most widely used modeling method to fit data is linear least squares regression. Least squares fitting (LSF) with a single variable of the number of carbon atoms in the *N*-*n*-alkyl chain, was first used in this study to build a simple linear relationship, i.e. Equation (2), to fit the affinity pK_i values of the 17 *N*-*n*-alkylnicotinium salts (Compounds 2–18 in Table I).

$$pK_i = 0.112x - 1.012 \quad (2)$$

The computational pK_i obtained from Equation (2) showed a linear relationship of the binding affinity (pK_i values) with the variation in the length of the *N*-*n*-alkyl chain. Assessment of the least-squares model indicated that the correlation coefficient r^2 of the regression from fitting was 0.46, which is less than 0.5 (Table II/Table IV); thus, this is not a statistically significant model.

Multiple linear regression analysis was then used to model the relationship between the pK_i values of the 17 *N*-*n*-alkylnicotinium salts with more explanatory variables. By searching the 143-descriptor variable space with a forward and reverse stepwise variable selection procedure, a linear model MLR1 (Equation (3)), was obtained:

$$pK_i = 0.532 + 6.083 \times E1m - 20.921 \times Gm \quad (3)$$

where E1m corresponds to the 1st component accessibility directional whim index / unweighted by atomic masses (the slash stands for divided by), and Gm corresponds to G total symmetry index / weighted by atomic masses, calculated by Dragon software [17]. The correlation coefficient between these two descriptors was -0.83 . The computed affinity of the

Table II. Quality of linear regression and ANN models.

Model	Training cycles	r^2	rmsd	r_{cv}^2	loormsd
LSF(17 mols)		0.46	0.69	0.33	0.77
MLR1(17 mols)		0.51	0.66	0.38	0.75
MLR2(32 mols)		0.79	0.48	0.71	0.57
MLR3(85 mols)		0.72	0.52	0.67	0.57
NN711(85 mols)	1000000	0.76	0.48	0.66	0.57
NN721(85 mols)	100000	0.82	0.42	0.68	0.56
NN731(85 mols)	80000	0.84	0.39	0.68	0.56
NN741(85 mols)	40000	0.82	0.42	0.66	0.57

N-*n*-alkylnicotinium salts for the $\alpha 4\beta 2^*$ nAChR subtype varied parabolically with increasing length of the *N*-*n*-alkyl chain (Figure 1b). A local minimum was found when the number of carbon atoms in the *n*-alkyl chain was equal to 4. The statistical analysis for the regression indicated that the correlation coefficient r^2 and rmsd between the observed and the fitted results was 0.51 and 0.66, respectively (Table II). The leave-one-out validation r_{cv}^2 was 0.38 and the leave-one-out rmsd was 0.75 (Table II). The Pearson correlation coefficients between calculated and observed pK_i values of the 17 molecules from fitting and LOO validation were 0.71 and 0.62 (Table IV), respectively.

To confirm the detected patterns from the above analysis, a multi-linear regression approach with a larger dataset was used to examine the relationship between the affinity of the *N*-*n*-alkylnicotinium salts at the $\alpha 4\beta 2^*$ nAChR subtype and increasing length of the *n*-alkyl chain. The dataset included 32 molecules (Table I) composed of 17 *N*-*n*-alkylnicotinium salts (Compounds 2–18 in Table I), 13 *N*-*n*-alkylpyridinium salts (Compounds 41–53 in Table I), and two *N*-*n*-alkylpicolinium salts (Compounds 65 and 66 in Table I). With a forward and reverse stepwise variable selection procedure, six descriptors E2u, Hy, LUMO, Me, Du and RDSUM, as defined in Table IIIa, were selected. A lower inter-correlation between these descriptors (non-diagonal element is larger than 0.85 in Table IIIb) indicated the possibility that the chance correlation for a linear equation built from these descriptors is low. A linear model (MLR2) between the six descriptors and the pK_i values of these 32 molecules was generated. Statistical analysis of the

Table IIIa. Brief description of the descriptors used the linear regression relationship (MLR2) with a 32-molecule dataset.

Descriptor	Definition
E2u	2nd component accessibility directional WHIM index / unweighted.
Hy	Hydrophilic factor.
LUMO	LUMO molecular orbital energy.
Me	Mean atomic Sanderson electronegativity (scaled on Carbon atom).
Du	D total accessibility index / unweighted.
RDSUM	Reciprocal distance Wiener-type index.

Table IIIb. Pearson correlation coefficient R between the descriptors used in MLR2 model.

	Eu	Hy	LUMO	Me	Du	RDSUM
Eu	1.00	0.21	-0.24	0.28	0.22	-0.18
Hy		1.00	-0.15	0.85	-0.49	-0.62
LUMO			1.00	-0.11	-0.20	0.78
Me				1.00	-0.36	-0.47
Du					1.00	-0.02
RDSUM						1.00

Table IV. Correlation between the observed pK_i values from the [^3H]nicotine binding assay and the calculated pK_i values for the series of 17 *N*-*n*-alkylnicotinium salts from various models.

Model	LSF	MLR1	MLR2	MLR3
R (Fitting)	0.68	0.71	0.73	0.70
Rmsd(Fitting)	0.69	0.66	0.65	0.73
R (LOO)	0.58	0.62	0.62	0.60
Rmsd(LOO)	0.77	0.75	0.75	0.77

Model	NN711	NN721	NN731	NN741
R (Fitting)	0.71	0.83	0.86	0.85
Rmsd(Fitting)	0.67	0.53	0.48	0.50
R (LOO)	0.59	0.61	0.68	0.65
Rmsd(LOO)	0.77	0.77	0.71	0.73

regression showed that the correlation coefficient r^2 was 0.79 with an rmsd of 0.48; leave-one-out validation r_{cv}^2 was 0.71 with an rmsd value of 0.57 (Table II). The linear equation is as follows:

$$pK_i = 130.228 - 2.729 \times \text{LUMO} + 0.071 \\ \times \text{RDSUM} + 41.877 \times \text{Hy} - 120.776 \times \text{Me} \\ - 6.022 \times \text{E2u} + 16.178 \times \text{Du} \quad (4)$$

The calculated pK_i values from Equation (4) for the 17 *N*-*n*-alkylnicotinium compounds *versus* the number of carbon atoms in the *N*-*n*-alkyl chain were plotted in Figure 1c. The fitting and validation Pearson correlation (R) between the experimental pK_i and computational pK_i values for the 17 *N*-*n*-alkylnicotinium molecules was 0.73 and 0.62, respectively (Table IV), which is similar to those obtained from MLR1. The computational results consistently showed that the affinity of the *N*-*n*-alkylnicotinium salts at the $\alpha 4\beta 2^*$ nAChR subtype parabolically varied with increasing length of the *N*-*n*-alkyl chain. A local minimum was found when the number of carbon in the *N*-*n*-alkyl chain was equal to four carbons atoms. Equation (4) was built by best fitting the affinity of 32 *mono*-nicotinium salts for the $\alpha 4\beta 2^*$ nAChR subtype, which helps reduce or eliminate the effects of experimental noise or error in analyzing the affinity of the *N*-*n*-alkylnicotinium salts for the $\alpha 4\beta 2^*$ nAChR subtype.

Results from neural network regression analysis

The artificial neural network (ANN) technique has been demonstrated recently to be an effective tool for data mining and has been used in many QSAR studies [19–23]. The major advantage of ANN lies in its ability to model a wide set of linear and non-linear functions, without knowing the analytic forms in advance. The ANN approach is especially suitable for mapping complex relationships that may exist between model inputs and output. To identify the best correlation between the observed and calculated pK_i values for the set of 85 compounds in the database we followed the same procedures as described previously for descriptor selection and determination of the optimal ANN configuration [18a,18b,24,25]. Variable selection from the dataset of 143 descriptors for the 85 molecules was carried out by a stepwise MLR procedure based on forward-selection and backward-elimination methods and located seven descriptors for an optimal multi-linear regression Equation (5) (MLR3). This afforded a correlation coefficient r^2 of 0.72 with an rmsd of 0.52; leave-one-out validation r_{cv}^2 of 0.67 with an rmsd value of 0.57 (Table II).

$$pK_i = 1.804 - 0.006 \times \text{PV} + 0.072 \times \text{PSA} \\ + 0.524 \times \text{LUMO} + 0.250 \times \text{DISNN} \\ - 0.165 \times \text{L2u} + 0.013 \times \text{RDSUM} \\ + 0.486 \times \text{MAXDP} \quad (5)$$

The inputs for the best ANN model were composed of the seven descriptors in Equation (5) i.e. polar volume (PV), polar surface area (PSA), lowest unoccupied molecular orbital energy (LUMO), distances between nitrogen atoms (DISNN), 2nd component size directional Whim index (L2u), reciprocal distance Wiener-type index (RDSUM), and maximal electrotopological positive variation (MAXDP), as indicated in Table Va. The correlation coefficient between the different descriptors is provided in Table Vb. No non-diagonal element was larger than 0.60 in Table Vb, indicating that redundant information included in the set of descriptors is low. Although the four configurations of the neural networks listed in Table II afforded close validation statistical results, the configurations NN721 and NN731 had a lower chance to be under-trained or over-trained for the 85 molecule dataset. The statistical results for model NN731 were as follows: $r^2 = 0.84$, rmsd = 0.39, $r_{cv}^2 = 0.68$, and loormsd = 0.56. The trained and LOO predicted versus observed pK_i values from NN731 were plotted in Figure 2.

The experimental and calculated pK_i values versus the number of carbon atoms in the *N*-*n*-alkyl chain of the 17 nicotinium salts are plotted in Figure 1e–h.

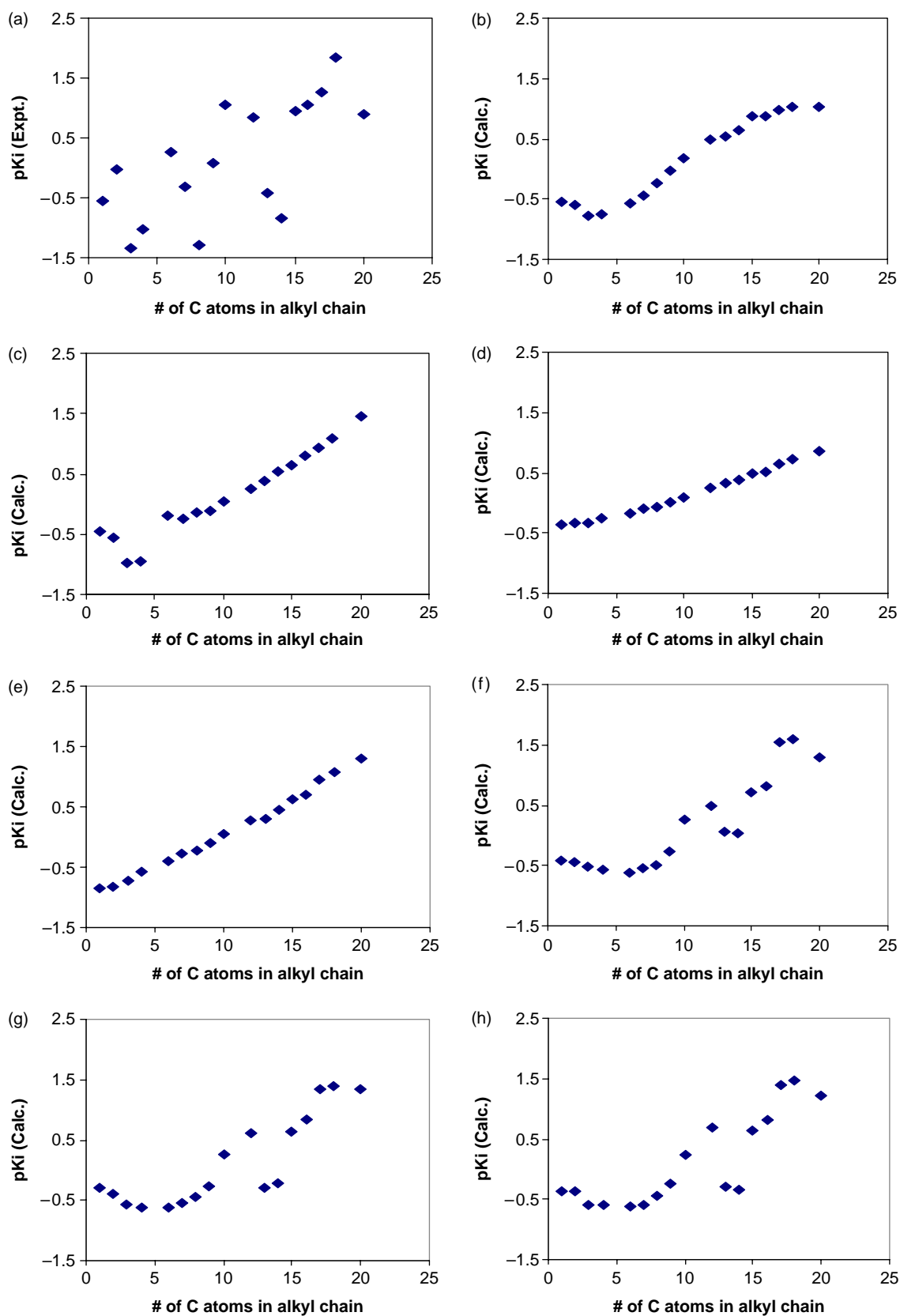


Figure 1. pK_i values versus the number of carbon atoms in the *N*-*n*-alkyl chain of the *N*-*n*-alkylnicotinium salts from: a. [^3H]nicotine binding assays; b. the MLR1 model with 17 molecules; c. the MLR2 model with 32 molecules; d. the MLR3 model with 85 molecules; e. the NN711 model; f. the NN721 model; g. the NN731 model; and h. the NN741 model with 85 molecules.

Table Va. Brief description of the descriptors used in the ANN models.

Descriptor	Definition
PV	Polar volume.
PSA	Polar surface area.
LUMO	LUMO molecular orbital energy.
DISNN	Distances between nitrogen atoms.
L2u	2nd component size directional Whim index.
RDSUM	Reciprocal distance Wiener-type index.
MAXDN	Maximal electrotopological positive variation.

Affinities of the *N*-*n*-alkylnicotinium salts predicted by NN711 (Figure 1e) show that increasing the length of the *N*-*n*-alkyl chain results in a linear increase in the affinity of the compounds for $\alpha 4\beta 2^*$ nAChRs, which is similar to Figure 1d (results obtained from MLR3). The lower correlation coefficient 0.71 for fitting and 0.59 for LOO validation between the computational and experimental pK_i values of the set of 17 homologous compounds (Table IV) indicated that the nature of the pattern among the set of data had not been fully identified. $\alpha 4\beta 2^*$ nAChR affinities of the *N*-*n*-alkylnicotinium salts predicted by both NN721 (Figure 1f) and NN731 (Figure 1g) showed that increasing the length of the *n*-alkyl chain resulted in a parabolic variation in affinity. A possible decrease in pK_i values may occur for some compounds with *n*-alkyl chain lengths ranging from C_{12} to C_{14} . The plot (Figure 1h) from NN741 had a very similar pattern to that from NN731. The pattern detected by NN731 afforded the best fit and validation Pearson correlation (0.86 and 0.68, respectively) with the observed data (Table IV). The pattern in Figure 1a [experimentally obtained affinity values (pK_i) versus the number of carbon atoms in the *n*-alkyl chain] was somewhat random, which is an interesting observation for further investigation. Generally, it is expected that the pK_i values would vary continuously, according to the basic chemical variation rule for a homologous series of compounds. The simplest interpretation for a decrease in the pK_i values for compounds with *n*-alkyl chain lengths in the range C_{12} to C_{13} from the pattern obtained from computational approaches, is that the binding site on the protein may be limited in size, and may not be able to accommodate the larger *N*-substituted molecules up to this point.

The subsequent increase in the pK_i values observed with molecules bearing *N*-*n*-alkyl substituents greater in length than C_{13} would suggest that these compounds with longer *n*-alkyl substituents bind to the receptor protein in a different manner, or at a different binding site. However, this remains to be proven in further studies.

Descriptor contributions to computational neural network models have been proposed and/or applied by several scientists [18, 24–27]. Utilizing the method proposed by Guha and Jurs [25], the results of the descriptor sensitivity analysis for NN731 showed that the percent contribution of the seven descriptors was: PV (14.3%), PSA (14.1%), LUMO (14.2%), DISNN (14.6%), L2u (13.9%), RDSUM (14.0%), and MAXDP (14.7%). The very similar percentage values for these seven descriptors suggest that the contributions of these descriptors to the model are almost equal. It should be noted that the above descriptor sensitivity analysis is related to a specific model. The parameters (weights) and transfer functions related to a single neuron in a neural network model also contribute to the power of the model's predictivity. In addition, a predictive model can be built from different sets of descriptors. Thus the descriptors having an important role in predicting target property are not limited only to the descriptors used to build the model. The correlation existing between the number of carbon atoms in an *n*-alkyl chain of a drug molecule containing a variable length *n*-alkyl substituent and the properties of these molecules, is well-known. These compounds, defined as members of a homologous series, have similar, as well as continuously varied properties. For example, a loss of potency as one ascends a homologous series of compounds is often used to map the dimensions of binding sites on a protein target [28]. Thus, whether these properties are modeled or not, the correlation is known to exist. The correlation between the number of carbon atoms in the alkyl chain of the *N*-*n*-alkylnicotinium salts and the properties of the molecule is another example of the above. With the small dataset of the *N*-*n*-alkylnicotinium salts utilized in this current study, one has only a limited appreciation of the relationship between the binding affinity and the length of the *n*-alkyl chain in these compounds. Neural network analysis can provide a much more comprehensive view of the

Table Vb. Pearson correlation coefficient R between the descriptors used in the ANN models.

	PV	PSA	LUMO	DISNN	L2u	RDSUM	MAXDP
PV	1.00	0.11	-0.27	0.40	-0.19	0.60	0.16
PSA		1.00	0.18	0.11	-0.02	-0.04	0.37
LUMO			1.00	0.53	0.25	-0.38	0.43
DISNN				1.00	0.34	0.37	0.49
W					1.00	0.33	0.11
RDSUM						1.00	0.11
MAXDP							1.00

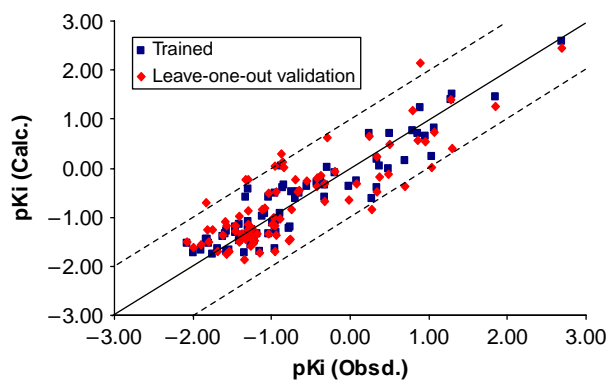


Figure 2. The calculated versus the experimentally determined pK_i values from the [^3H]nicotine binding assay for the trained (shown in black squares), leave-one-out cross-validation (shown in red diamonds) for the best NN731 QSAR model. The solid line represents a perfect correlation. The dotted lines represent one order difference from the perfect fitting. Most points around the dotted lines correspond to the N - n -alkylnicotinium compounds in Table VI.

structure-activity relationships because of its ability to model a wide set of linear and non-linear functions, without knowing the analytic forms in advance. However, with a small number of compounds, the full predictive power of neural network analysis cannot be utilized, due to possible overtraining problems. Generally, when a dataset includes less than 40 compounds, back propagation neural network analysis will just model patterns that are not much different from linear function modeling. In this current study, we built more predictive models to detect patterns from a larger number of 85 compounds, by utilizing the advantages of the neural network approach and, in a creative way, graphically expressed the relationship of the binding affinity of the N - n -alkylnicotinium salts with the number of carbon atom in the n -alkyl chain from the neural network modeling results. Obviously, whereas a linear relationship can only detect major features, such as the longer the alkyl chain, the larger the pK_i value of a compound (Figure 1d,e). Figure 1b–c,f–h with a higher coefficient R value and smaller RMSD (Table IV) show the affinity values (pK_i) vary

near-parabolically with increasing alkyl chain length with a local minimum at C4. From Table IV, one can also see that a higher R (Fitting) corresponds to a better R (LOO), which makes the results more reliable.

Table VI indicates that 10 nicotinium analogs from the set of 85 *mono*- and *bis*-quaternary ammonium salts had absolute error values between the experimental and calculated pK_i values of larger than 0.80, either in training or leave-one-out, or in both, when comparing the observed and calculated pK_i values in model NN731. The original training correlation coefficient r^2 of the 85 molecule dataset from the neural network model NN731 was 0.84 with an rmsd of 0.39; leave-one-out r_{cv}^2 was 0.68 with an rmsd of 0.56. After removing the 10 compounds listed in Table VI from the trained model, the statistical analysis showed that the training correlation coefficient r^2 for the remaining data was 0.92 with an rmsd of 0.25; leave-one-out r_{cv}^2 was 0.80 with an rmsd of 0.41. Apparently, the difficulty in obtaining a better model from the dataset of 85 compounds is mainly due to the error associated with the experimental data from the nicotinium series of compounds. The same conclusion holds in the analysis of the data obtained from model NN721.

Comparison with the previously published computational models

As an important target for drug development, several modeling approaches have been carried out in order to understand quantitative structure-activity relationships of ligands that bind to nAChR subtypes. Two recent reviews that report on ligand-based models of neuronal nAChR agonists, involving either pharmacophore development or quantitative structure-affinity relationships, have been reported by Glennon et al. and Nicolotti et al., respectively [29–30]. More recently, QSAR studies in this area that focus on specific types of analogs, have also been reported [31–35]. With regard to quaternary ammonium nAChR antagonists, QSAR studies on a series of *mono*- and *bis*-quaternary ammonium salts have been reported by Ayers et al.

Table VI. Compounds with large errors between the experimentally determined pK_i values from [^3H]nicotine binding assays and calculated pK_i /LOO pK_i values by model NN731.

	pK_i (expt.)	pK_i (NN731)	Diff.	pK_i (NN731LOO)	Diff.
NONB-3C	1.10	-0.56	1.66	-0.77	1.86
NDNB-4C	-0.87	0.05	-0.92	0.29	-1.16
NONB-6e7m	0.35	-0.39	0.74	-0.49	0.83
NHxNI	0.28	-0.61	0.89	-0.84	1.12
GZ-511B	-0.85	-0.35	-0.50	0.01	-0.85
NPNI	-1.33	-0.59	-0.74	-0.25	-1.09
NDNI	1.05	0.23	0.82	0.01	1.04
NONI	-1.29	-0.44	-0.86	-0.24	-1.05
NONB-3Y	0.70	0.14	0.55	-0.39	1.08
NCYNB-6	-0.86	-0.41	-0.45	0.13	-0.99

[36]. In the study of Ayers et al, the Self Organizing Maps (SOM) approach was used to classify the compounds according to their structures and activities. Several genetic functional approximation models were created to simulate the quantitative structure-activity relationships from three small subsets of compound datasets (i.e., sets of 31, 38 and 23 compounds).

Our current work focuses on the pattern detection for the affinity of *N*-*n*-alkyl quaternary ammonium salts in the [^3H]nicotine binding assay using rat striatal membranes *versus* the number of carbon atoms in the *N*-*n*-alkyl substituent by mining knowledge from 17, 32 and 85 molecule datasets. Models generated from this larger 85-molecule dataset may provide additional power for future prediction of K_i values of new virtual compounds.

Conclusions

Linear regression and neural network approaches have been used to build mathematical relationships to compute the binding affinity of a series of *mono*- and *bis*-quaternary ammonium salts in [^3H]nicotine binding assays using rat striatal membranes. These results were then used to analyze the pattern among the experimentally determined affinities (K_i values) of a set of 17 *N*-*n*-alkylnicotinium salts for the $\alpha 4\beta 2^*$ nicotinic receptor subtype. The affinity of these *N*-*n*-alkylnicotinium compounds was shown to vary parabolically with increasing length of the *n*-alkyl chain, with a local minimum indicated for the C_4 analogue. A decrease in $\text{p}K_i$ values for compounds with *n*-alkyl chain lengths of C_{12} and C_{13} was also evident. The generated neural network model with the larger 85 molecule training set may provide additional power for future prediction of K_i values of new virtual compounds, prior to their synthesis and pharmacological evaluation.

Acknowledgement

This work was supported by NIH grant U19DA017548.

References

- [1] Quik M. Trends Neurosci 2004;27:561–568.
- [2] Hogg RC, Bertrand D. Science 2004;306:983–985.
- [3] Tapper AR, McKinney SL, Nashmi R, Schwarz J, Deshpande P, Labarca C, Whiteaker P, Marks MJ, Collins AC, Lester HA. Science 2004;306:1029–1032.
- [4] Paradiso KG, Steinbach JH. J Physiol 2003;553:857–871.
- [5] Jensen AA, Frolund B, Liljefors T, Krogsgaard-Larsen P. J Med Chem 2005;48(15):4705–4745.
- [6] Romanelli MN, Gualtieri F. Med Res Rev 2003;23(4):393–397.
- [7] (a) Crooks PA, Ayers JT, Xu R, Sumithran SP, Grinevich VP, Wilkins LH, Deaciuc AG, Allen DD, Dwoskin LP. Bioorg Med Chem Letters 2004;14:1869–1874. (b) Grinevich VP, Crooks PA, Haubner AJ, Ghosheh OH, Ayers JH, Dwoskin LP. J Pharmacol Exp Ther 2003; 306:1011–1020. (c) Ayers JT, Dwoskin LP, Deaciuc AG, Grinevich VP, Zhu J, Crooks PA. Bioorg Med Chem Lett 2002; 12:3067–3071.
- [8] Dwoskin LP, Sumithran SP, Zhu J, Deaciuc AG, Ayers JT, Crooks PA. Bioorg Med Chem Letters 2004;14:1863–1865.
- [9] Xu R, Dwoskin LP, Grinevich V, Sumithran SP, Crooks PA. Drug Dev Res 2002;55:173–186.
- [10] Wilkins LH, Haubner A, Ayers JT, Crooks PA, Dwoskin LP. J Pharmacol Exp Ther 2002;301:1088–1096.
- [11] Wilkins LH, Grinevich VP, Ayers JT, Crooks PA, Dwoskin LP. J Pharmacol Exp Ther 2003;304:400–410.
- [12] Crooks PA, Ravard A, Wilkins LH, Teng LH, Buxton ST, Dwoskin LP. Drug Dev Res 1995;36:91–102.
- [13] Dwoskin LP, Crooks PA. J Pharmacol Exp Ther 2001;298:395–402.
- [14] The $\text{p}K_a$ value was calculated according to the same approach described in (a) Huang XQ, Zheng F, Crooks PA, Dwoskin LP, Zhan C-G. J Am Chem Soc 2005;127:14401–14414. (b) Huang XQ, Zheng F, Chen X, Crooks P A, Dwoskin LP, Zhan C-G. J Med Chem 2006; 49(26):7661–7674.
- [15] Edelstein HA. Introduction to data mining and knowledge discovery. 3rd ed. Potomac, MD: Two Crows Corp; 1999.
- [16] (a) Tripos discovery software package with SYBYL 6.8.1, Tripos Inc., 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA. (b) Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery Jr JA, Vrener T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham M, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA. Gaussian 03, Revision A.1, Gaussian, Inc., Pittsburgh, PA, 2003.
- [17] DRAGON software version 3.0, developed by Milano Chemometrics and QSAR Research Group, 2003 (<http://www.disat.nimib.it/chm/Dragon.htm>).
- [18] (a) Zheng F, Bayram E, Sumithran SP, Ayers JT, Zhan C-G, Schmitt JD, Dwoskin LP, Crooks PA. Bioorg Med Chem 2006;14(9):3017–3037. (b) Zheng F, Zheng G, Deaciuc AG, Zhan C-G, Dwoskin LP, Crooks PA. Bioorg Med Chem 2007; 15:2975-2992.
- [19] (a) Orre R, Lansner AB, Lindquist M. Computational statistics and data analysis 2000;34:473–493. (b) Lancashire LJ, Mian S, Ellis IO, Rees RC, Ball GR. Current Proteomics 2005; 2(1):15-29.
- [20] Fernandez M, Caballero J. J Mol Graph Model 2006;25:410–422.
- [21] Fernández M, Carreiras MC, Marco JL, Caballero J. J Enz Inhib Med Chem 2006;21:647–661.
- [22] Caballero J, Tundidor-Camba A, Fernández M. QSAR Comb Sci 2007;26:27–40.
- [23] González MP, Caballero J, Tundidor-Camba A, Helguera AM, Fernández M. Bioorg Med Chem 2006;14:200–213.
- [24] Guha R, Jurs PC. J Chem Inf Model 2005;45:800–806.
- [25] Guha R, Stanton DT, Jurs PC. J Chem Inf Model 2005;45:1109–1121.
- [26] Caballero J, Fernandez L, Abreu JI, Fernandez M. J Chem Inf Model 2006;46:1255–1268.
- [27] Cherqaoui D, Villemin D. J Chem Soc Faraday Trans 1994; 90:97–102.

- [28] Eckenhoff RG, Tanner JW, Johansson JS. *Mol Pharmacol* 1999;56:414–418.
- [29] Glennon RA, Dukat M, Liao L. *Curr Top Med Chem* 2004; 4(6):631–644.
- [30] Nicolotti O, Altomare C, Pellegrini-Calace M, Carotti A. *Curr Top Med Chem* 2004;4(3):335–360.
- [31] Zhang H, Li H, Liu C. *J Chem Inf Model* 2005;45(2):440–448.
- [32] Zhang H, Li H, Ma Q. *J Mol Graph Model* 2007;26(1):226–235.
- [33] Audouze K, Nielsen EØ, Olsen GM, Ahring P, Jørgensen TD, Peters D, Liljefors T, Balle T. *J Med Chem* 2006;49(11): 3159–3171.
- [34] Ji J, Schrimpf MR, Sippy KB, Bunnelle WH, Li T, Anderson DJ, Faltynek C, Surowy CS, Dyhring T, Ahring PK, Meyer MD. *J Med Chem* 2007;50(22):5493–5508.
- [35] Basu A, Gayen S, Samanta S, Panda P, Srikanth K, Jha T. *Canadian J Chem* 2006;84(3):458–463.
- [36] Ayers JT, Clauser A, Schmitt JD, Dwoskin LP, Crooks PA. *AAPS Journal: AAPS-NIDA Symposium, Frontiers in Science: Drug Addiction: From Basic Research to Therapeutics* 2005;7(3):E678–E685.