Feature Selection and Linear/Nonlinear Regression Methods for the Accurate Prediction of Glycogen Synthase Kinase-3β Inhibitory Activities

Mohammad Goodarzi,[†,‡] Matheus P. Freitas,*[,||] and Richard Jensen[§]

Department of Chemistry, Faculty of Sciences, Azad University, Arak, Iran, Young Researchers Club, Azad University, Arak, Iran, Department of Computer Science, The University of Wales, Aberystwyth, U.K., and Departamento de Química, Universidade Federal de Lavras, CP 3037, 37200-000, Lavras, MG, Brazil

Received January 12, 2009

Few variables were selected from a pool of calculated Dragon descriptors through three different feature selection methods, namely genetic algorithm (GA), successive projections algorithm (SPA), and fuzzy rough set ant colony optimization (fuzzy rough set ACO). Each set of selected descriptors was regressed against the bioactivities of a series of glycogen synthase kinase-3β (GSK-3β) inhibitors, through linear and nonlinear regression methods, namely multiple linear regression (MLR), artificial neural network (ANN), and support vector machines (SVM). The fuzzy rough set ACO/SVM-based model gave the best estimation/prediction results, demonstrating the nonlinear nature of this analysis and suggesting fuzzy rough set ACO, first introduced in chemistry here, as an improved variable selection method in QSAR for the class of GSK-3β inhibitors.

1. INTRODUCTION

Many QSAR strategies have been introduced in chemistry to model the bioactivities of a variety of druglike compounds. In addition to receptor-based approaches, structure-based QSAR methodologies have given highly predictive models, and most of them use three-dimensional descriptors[1−3] or multidimensional inferences.[4−6] Similarly, image-based analysis has also provided useful descriptors in QSAR.[7−9] However, classical and/or 2D structural descriptors have shown to be not inferior to 3D/*n*D ones, at least in many practical cases.[10] This has been improved with the development of several algorithms for feature selection and regression between dependent and independent variables.

Multiple linear regression (MLR) and, for a larger number of descriptors, partial least-squares (PLS) are usually applied in correlation. However, methods like these do not account for nonlinearity and, when used without any variable selection method, may offer poorly predictive models. In line with this, a variety of feature selection methods have been developed to select suitable QSAR descriptors. Genetic algorithms (GAs) arise as a method of widespread use for dimension reduction and have been reported in the optimization of a number of different and traditionally difficult problems, including image processing, design of complex networks (e.g., computers and integrated circuits), classifications, parameters for neural nets, job scheduling, robotics, and parameter fitting.[11] Additionally, successive projections algorithm (SPA) has been more recently applied to solve spectral mixing problems;[12] the SPA algorithm extracts endmembers from hyperspectral data without having to reduce the data dimensionality, but it is not limited to deal with spectral information. Furthermore, in this work, we have coupled fuzzy lower approximation, rough set-based feature selection, and ant colony optimization (ACO).

Ant colony optimization (ACO)[13] is an area of interest within swarm intelligence. It can be understood as real ants that are capable of finding the shortest route between a food source and their nest without the use of visual information and hence possess no global world model, adapting to changes in the environment. The deposition of pheromone is the main factor in enabling real ants to find the shortest routes over a period of time. Each ant probabilistically prefers to follow a direction rich in this chemical. The pheromone decays over time, resulting in much less pheromone on less popular paths. Given that over time the shortest route will have the higher rate of ant traversal, this path will be reinforced and the others diminished until all ants follow the same, shortest path (the "system" has converged to a single solution). It is also possible that there are many equally short paths. In this situation, the rates of ant traversal over the short paths will be roughly the same, resulting in these paths being maintained while others are ignored. Additionally, if a sudden change to the environment occurs (e.g., a large obstacle appears on the shortest path), the ACO system can respond to this and will eventually converge to a new solution. Based on this idea, artificial ants can be deployed to solve complex optimization problems via the use of artificial pheromone deposition. ACO is particularly attractive for feature selection as there seems to be no heuristic that can guide search to the optimal minimal subset every time. Additionally, it can be the case that ants discover the best feature combinations as they proceed throughout the search space.

The feature selection task may be reformulated into an ACO-suitable problem.[14,15] ACO requires a problem to be represented as a graph - nodes represent features, with the

* Corresponding author phone: +55 35 3829-1891; fax: +55 35 3829-1271; e-mail: matheus@ufla.br.
† Faculty of Sciences, Azad University.
‡ Young Researchers Club, Azad University.
§ The University of Wales.
|| Universidade Federal de Lavras.

10.1021/ci9000103 CCC: $40.75
Published on Web 04/01/2009

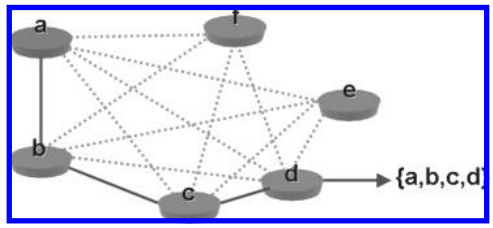


Figure 1. ACO problem representation for feature selection.

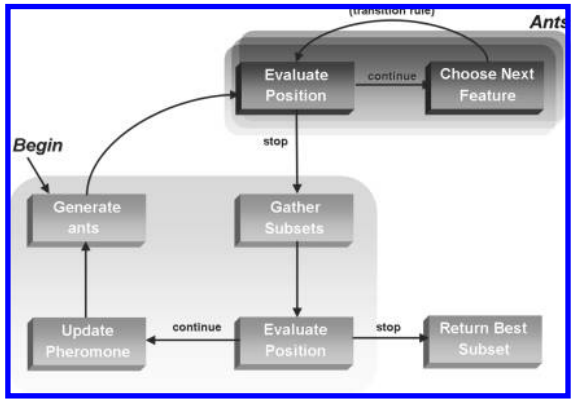


Figure 2. ACO-based feature selection overview.

Table 1. Comparison of Estimations Using the Original **Y** Block and the Validation Procedure Using the Randomized **Y** Block

		FRS-ACO	GA	SPA
number of trials		50	50	50
r^2 for original Y		0.806	0.770	0.773
r^2 for randomized-Y	avg	0.097	0.112	0.099
	max	0.211	0.197	0.167
	SD	0.006	0.009	0.005

edges between them denoting the choice of the next feature (Figure 1). The search for the optimal feature subset is then an ant traversal through the graph where a minimum number of nodes are visited that satisfies the traversal stopping criterion.

A suitable heuristic desirability of traversing between features could be any subset evaluation function; for example, an entropy-based measure or the fuzzy-rough set dependency measure. Depending on how optimality is defined for the particular application, the pheromone may be updated accordingly. For instance, subset minimality and "goodness" are two key factors so the pheromone update should be proportional to "goodness" and inversely proportional to size. How "goodness" is determined will also depend on the application. In some cases, this may be a heuristic evaluation of the subset; in others it may be based on the resulting classification accuracy of a classifier produced using the subset.

The heuristic desirability and pheromone factors are combined to form the probabilistic transition rule, denoting the probability of an ant k at feature i choosing to move to feature j at time t

$$p_{i,j}^{k}(t) = \frac{[\tau_{i,j}(t)]^{\alpha} \cdot [\eta_{i,j}]^{\beta}}{\sum_{l \in J_i^k} [\tau_{i,l}(t)]^{\alpha} \cdot [\eta_{i,l}]^{\beta}} \quad (1)$$

where J_i^k is the set of ant k's unvisited features, $\eta_{i,j}$ is the heuristic desirability of choosing feature j when at feature i, and $\tau_{i,j}(t)$ is the amount of virtual pheromone on edge (i,j). The choice of α and β is determined experimentally.

The overall process of ACO feature selection can be seen in Figure 2. It begins by generating a number of ants, k, which are then placed randomly on the graph (i.e., each ant starts with one random feature). Alternatively, the number of ants to place on the graph may be set equal to the number of features within the data; each ant starts path construction at a different feature. From these initial positions, they traverse edges probabilistically until a traversal stopping criterion is satisfied. The resulting subsets are gathered and then evaluated. If an optimal subset has been found or the algorithm has executed a certain number of times, then the process halts and outputs the best feature subset encountered. If neither condition holds, then the pheromone is updated, a new set of ants are created, and the process iterates once more.

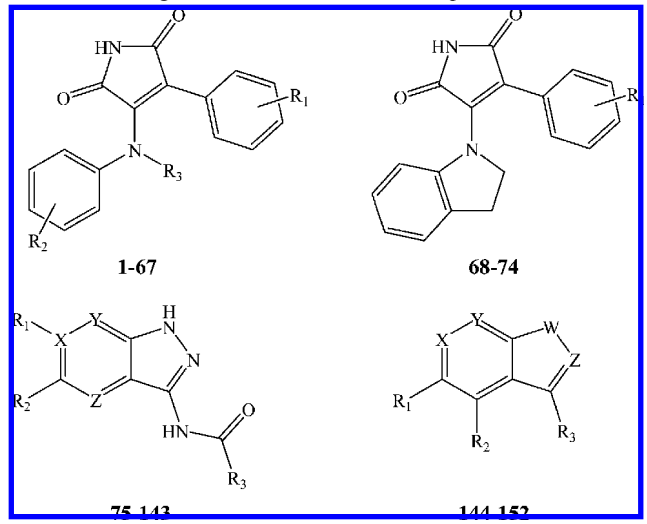
Rough set theory (RST) can be used as a tool to discover data dependencies and to reduce the number of attributes contained in a data set using the data alone, requiring no additional information.[16,17] The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same predictive capability of the decision feature as the original. This rough set attribute reduction (RSAR) process[18,19] can only operate effectively with data sets containing discrete values. Additionally, there is no way of handling noisy data. As most data sets contain real-valued attributes, it is necessary to perform a discretization step beforehand, often resulting in significant information loss. A way of handling this problem is through the use of fuzzy-rough sets. Subjective judgments are not entirely removed as fuzzy set membership functions still need to be defined. However, the method offers a high degree of flexibility when dealing with real-valued data, enabling the vagueness and imprecision present to be modeled effectively. Thus, together with ACO, fuzzy rough sets are supposed to be a powerful tool to select suitable descriptors and improve QSAR models.

Together with the application of suitable variable selection methods, the regression step leads to the success of a QSAR modeling. While methods like MLR do not account for nonlinear behavior, artificial neural networks (ANN), such as optimized backpropagation ANN, estimate the global error vector prior to adjusting weights and update successively the weights until convergence is reached. Optimization starts with a backpropagation function which is solved by the nonlinear Levenberg−Marquardt algorithm.[20] Moreover, support vector machines (SVM) is a relatively new nonlinear technique in the field of chemometrics and is employed in classification and multivariate calibration problems.[21−23] SVM is capable of dealing with linear and nonlinear multivariate calibration and resolves multivariate calibration problems in a relatively fast way. Therefore, the joint use of prominent feature selection methods and regression algorithms on easily obtained descriptors to a series of glycogen synthase kinase-3β (GSK-3β) inhibitors was applied here to give predictable QSAR models.

2. COMPUTATIONAL METHODS

The 2D structures of 132 training set and 29 external test compounds obtained from the literature[24] were drawn using

Table 2. Series of GSK-3β Inhibitors Used in Training and Test Sets[a] and Their Respective Bioactivities (in nM)



cpd	R1	R2	R3	X	Y	W	Z	IC_{50}
1^{+}	H	H	H					590
2	H	3-chloro	H					301
3	H	3-hydroxy	H					704
4*	H	3,5-dichloro-4-hydroxy	H					149
5	H	3-carboxy	H					291
6	H	4-chloro-3-carboxy	H					143
7^{+}	H	4-SCH_3	H					404
8	2-chloro	H	H					216
9*	2-chloro	3-chloro	H					195
10	2-chloro	3-hydroxy	H					374
11^{+}	2-chloro	3-chloro-4-hydroxy	H					152
12	2-chloro	3,5-dichloro-4-hydroxy	H					93
13	2-chloro	3-carboxy	H					136
14^{+}	2-chloro	4-chloro-3-carboxy	H					74
15	2-chloro	4-SCH_3	H					161
16*	2-methoxy	H	H					216
17	2-methoxy	3-chloro	H					114
18^{+}	2-methoxy	3-hydroxy	H					259
19*	2-methoxy	3-chloro-4-hydroxy	H					139
20	2-methoxy	3,5-dichloro-4-hydroxy	H					82
21	2-methoxy	4-SCH_3	H					110
22*	2-nitro	3-chloro	H					104
23	2-nitro	3-hydroxy	H					251
24^{+}	2-nitro	3-chloro-4-hydroxy	H					104
25^{+}	2-nitro	3,5-dichloro-4-hydroxy	H					52
26	2-nitro	4-chloro-3-carboxy	H					28
27^{+}	3-chloro	3-hydroxy	H					1478
28^{+}	3-chloro	3-chloro-4-hydroxy	H					94
29	3-chloro	3,5-dichloro-4-hydroxy	H					58
30	3-chloro	3-carboxy	H					134
31	3-chloro	4-chloro-3-carboxy	H					76
32	3-chloro	4-SCH_3	H					532
33	3-methoxy	3-chloro	H					257
34*	3-methoxy	3-hydroxy	H					472
35	3-methoxy	3,5-dichloro-4-hydroxy	H					142
36	3-methoxy	3-carboxy	H					195
37	3-methoxy	4-chloro-3-carboxy	H					85
38^{+}	3-methoxy	4-SCH_3	H					203
39^{+}	3-nitro	H	H					141
40*	3-nitro	3-chloro	H					70
41	3-nitro	3-hydroxy	H					236
42*	3-nitro	4-hydroxy	H					123
43	3-nitro	3-chloro-4-hydroxy	H					59
44^{+}	3-nitro	3,5-dichloro-4-hydroxy	H					20
45*	3-nitro	3-carboxy	H					79
46	3-nitro	4-chloro-3-carboxy	H					26
47	3-nitro	4-SCH_3	H					152
48	4-chloro	H	H					514
49	4-chloro	3-chloro	H					447
50	4-chloro	3-hydroxy	H					407
51^{+}	4-chloro	4-hydroxy	H					317

Table 2. Continued

cpd	R1	R2	R3	X	Y	W	Z	IC_{50}
52	4-chloro	3-chloro-4-hydroxy	H					173
53*	4-chloro	3,5-dichloro-4-hydroxy	H					91
54	4-chloro	3-carboxy	H					186
55	4-chloro	4-chloro-3-carboxy	H					109
56*	4-chloro	4-SCH_3	H					529
57*	4-methoxy	H	H					390
58$^+$	4-methoxy	3-chloro	H					156
59	4-methoxy	3-hydroxy	H					481
60	4-methoxy	3,5-dichloro-4-hydroxy	H					83
61	4-methoxy	3-carboxy	H					214
62*	4-methoxy	4-SCH_3	H					243
63	4-nitro	3,5-dichloro-4-hydroxy	H					71
64	4-nitro	4-SCH_3	H					392
65	H	H	methyl					2613
66$^+$	3-nitro	H	methyl					1398
67*	4-chloro	H	methyl					2285
68	2-chloro							337
69	2-methoxy							187
70	2-nitro							131
71$^+$	3-chloro							460
72*	3-nitro							161
73	4-chloro							1412
74*	4-methoxy							694
75		phenyl	*n*-propyl	CH	CH		CH	99
76		phenyl	*n*-propyl	N	CH		CH	7
77		phenyl	*n*-propyl	N	CH		N	2697
78		phenyl	*n*-propyl	N	N		C-phenyl	691
79*		phenyl	$(CH_2)_3N(CH_3)_2$	N	N		CH	22
80		phenyl	$(CH_2)_3$-pyrrolidine	N	N		CH	11
81$^+$		phenyl	$(CH_2)_3$-piperazinyl-*N*-(C_2H_5)	N	N		CH	7
82		phenyl	$(CH_2)_3$-morpholinyl	N	N		CH	5
83		phenyl	4-piperidine-NCH_3	N	N		CH	9
84$^+$		phenyl	CH_2-4-piperazinyl-*N*-(C_2H_5)	N	N		CH	5
85		2,2-difluoro-phenyl	$(CH_2)_3N(CH_3)_2$	N	N		CH	5
86	phenyl	bromo	cyclopropyl	CH	N		CH	75
87	4-OH-phenyl	bromo	cyclopropyl	CH	N		CH	0.8
88	4-OH-phenyl	H	cyclopropyl	CH	N		CH	8
89*	3-bromo-4-OH- phenyl	H	cyclopropyl	CH	N		CH	5
90*	3-chloro-4-OH- phenyl	H	cyclopropyl	CH	N		CH	7
91	4-OH-phenyl	phenyl	cyclopropyl	CH	N		CH	24
92$^+$	4-OH-phenyl	bromo	$(CH_2)_3$-4-piperazinyl-*N*-(C_2H_5)	CH	N		CH	4
93	3-OH-phenyl	H	cyclopropyl	CH	N		CH	12
94	2-thienyl	bromo	cyclopropyl	CH	N		CH	39
95	2-thienyl	bromo	cyclopropyl	CH	N		CH	7
96	2-furyl	H	cyclopropyl	CH	N		CH	141
97	2-furyl	bromo	cyclopropyl	CH	N		CH	7
98	2-thiazoyl	bromo	cyclopropyl	CH	N		CH	99
99	2-thiazoyl	bromo	cyclopropyl	CH	N		CH	16
100	2-thienyl	bromo	CH_2-4-piperidine-*N*-CH_3	CH	N		CH	18
101	2-furyl	bromo	3-pyrrolidine-*N*-benzyl	CH	N		CH	14
102	H	phenyl	methyl	CH	N		CH	291
103	H	phenyl	ethyl	CH	N		CH	43
104	H	phenyl	*n*-propyl	CH	N		CH	56
105	H	phenyl	isopropyl	CH	N		CH	19
106	H	phenyl	cyclopentyl	CH	N		CH	5
107	H	phenyl	$NH(C_2H_5)$	CH	N		CH	2810
108	H	$B(OC(CH_3)_2$-$C(CH_3)_2O)$	*n*-propyl	CH	N		CH	356
109	H	H	*n*-propyl	CH	N		CH	2343
110*	H	2-fluoro-phenyl	*n*-propyl	CH	N		CH	18
111	H	3- fluoro-phenyl	*n*-propyl	CH	N		CH	20
112	H	2,3-difluoro-phenyl	*n*-propyl	CH	N		CH	7
113	H	2-chloro-phenyl	*n*-propyl	CH	N		CH	27
114	H	3-pyridyl	*n*-propyl	CH	N		CH	11
115	H	4-pyridyl	*n*-propyl	CH	N		CH	443
116	H	4-bisphenyl	*n*-propyl	CH	N		CH	851
117	H	2-naphthyl	*n*-propyl	CH	N		CH	169
118	H	1-naphthyl	*n*-propyl	CH	N		CH	241
119*	phenyl	H	cyclopropyl	CH	N		CH	425
120*	3,4-di-OH-phenyl	H	cyclopropyl	CH	N		CH	8
121	3-OCH_3-phenyl	H	cyclopropyl	CH	N		CH	125
122	2-OH-phenyl	H	cyclopropyl	CH	N		CH	36
123$^+$	2-OCH_3-phenyl	H	cyclopropyl	CH	N		CH	1593

Table 2. Continued

cpd	R1	R2	R3	X	Y	W	Z	IC_{50}
124^{+}	4-OH-phenyl	chloro	cyclopropyl	CH	N		CH	1
125	4-OH-phenyl	methyl	cyclopropyl	CH	N		CH	6
126	phenyl	phenyl	cyclopropyl	CH	N		CH	415
127	phenyl	chloro	cyclopropyl	CH	N		CH	234
128	phenyl	CN	cyclopropyl	CH	N		CH	87
129*	phenyl	bromo	4-piperidine-*N*-(CH_3)	CH	N		CH	383
130*	4-OH-phenyl	H	4-piperidine-*N*-(CH_3)	CH	N		CH	12
131	4-OH-phenyl	bromo	4-piperidine-*N*-(CH_3)	CH	N		CH	1
132	3-OH-phenyl	H	(CH_2)-3-piperazinyl-*N*-(C_2H_5)	CH	N		CH	21
133^{+}	4-OCH_3-phenyl	H	cyclopropyl	CH	N		CH	23000
134	phenyl		cyclopropyl	CH	CH		CH	498
135	4-OH-phenyl		cyclopropyl	CH	CH		CH	15
136^{+}	5-indolyl		cyclopropyl	CH	CH		CH	42
137	phenyl-3-SO_2NH_2		cyclopropyl	CH	CH		CH	481
138	3-fluoro- phenyl		cyclopropyl	CH	CH		CH	828
139*	2-pyrrolyl		cyclopropyl	CH	CH		CH	320
140*	3-furanyl		cyclopropyl	CH	CH		CH	35
141	2-thienyl		cyclopropyl	CH	CH		CH	215
142	3-thienyl		cyclopropyl	CH	CH		CH	329
143	2,5-difluoro-phenyl		cyclopropyl	CH	CH		CH	1000
144	phenyl	phenyl	NH_2	N	N	NH	N	250
145	phenyl	H	NH_2	N	N	NH	N	530
146	phenyl	H	NH_2	CH	N	NH	N	430
147*	phenyl	H	NH_2	N	CH	NH	N	1260
148	H	phenyl	NH_2	CH	N	NH	N	23000
149*	phenyl		$NHSO_2(CH_3)$	CH	N	NH	N	3572
150	phenyl		NH_2	CH	N	N(CH_3)	N	23000
151	phenyl		NH_2	CH	N	O	N	23000
152*	phenyl		NH_2	CH	N	NH	CH	23000

[a] Compounds with an asterisk pertain to the test set. Compounds with a + pertain to the validation set in the ANN model.

Table 3. Dragon Descriptors Selected by Genetic Algorithm (GA)

symbol	definition	class
Mor01p	3D-MoRSE-signal 01/weighted by atomic polarizabilities	3D MoRSE descriptors
GATS8m	Geary autocorrelation lag 8/weighted by atomic masses	2D autocorrelation
MATS3p	Moran autocorrelation-lag3/weighted by atomic polarizabilities	2D autocorrelation
Se	sum of atomic Sanderson electronegativities (scaled on carbon atom)	constitutional descriptors
GGI3	topological charge index of order 3	Galvez topol, charge indices
JGI1	mean topological charge index of order 1	Galvez topol, charge indices
TIC2	total information content index (neighborhood symmetry of 2-order)	topological descriptors
MATS6v	Moran autocorrelation lag-6/weighted by atomic van der Waals volumes	2D autocorrelation
Mor06e	3D-MoRSE-signal 6/weighted by atomic Sanderson electronegativities	3D MoRSE descriptors
Ms	mean electrotopological state	constitutional descriptors
qpmax	maximum positive charge	charge descriptors
R5v+	R matrix autocorrelation lag5/weighted by atomic van der Waals volumes	GETAWAY descriptors
RDF025e	radial distribution function-2.5/weighted by atomic Sanderson elecronegativities	RDF descriptors
R6u+	R maximal autocorrelation of lag6/unweighted	GETAWAY descriptors
MWC02	molecular walk count of order 02	molecular walk counts
ZM1V	first Zagreb index by valence vertex degree	topological descriptors

Hyperchem 7 software.[25] Geometries were preoptimized with the Molecular Mechanics Force Field (MM+), and the final ones were obtained with the semiempirical AM1 method in the Hyperchem program. The molecular structures were optimized using the Polak-Ribiere algorithm until the root-mean-square gradient was 0.001 kcal mol^{-1}. The resulting geometry was transferred into the Dragon program package,[26] in order to calculate about 1457 descriptors of constitutional, topological, geometrical, charge, GETAWAY (Geometry, Topology and Atoms-Weighted Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), molecular walk count, BCUT, 2D-autocorrelation, aromaticity index, Randic molecular profile, radial distribution function, functional group, and atom-centered fragment classes. It is worth mentioning that all descriptors with the same values for all molecules were omitted, and also one of the two descriptors that has the pair wise correlation coefficient above 0.9 ($r > 0.9$) and has a large correlation coefficient with the other descriptors in each class was eliminated. By the way, to demonstrate the absence of chance correlations on the best models obtained with the

Table 4. Dragon Descriptors Selected by Successive Projections Algorithm (SPA)

symbol	definition	class
Mor20m	3D-MoRSE-signal 20/weighted by atomic masses	3D MoRSE descriptors
P1p	1st component shape directional WHIM index/weighted by atomic polarizabilities	WHIM descriptors
Sv	sum of atomic van der Waals volumes (scaled on carbon atom)	constitutional descriptors
Se	sum of atomic Sanderson electronegativities (scaled on carbon atom)	constitutional descriptors
Sp	sum of atomic polarizabilities (scaled on carbon atom)	constitutional descriptors
G2u	2nd component symmetry directional WHIM index/unweighted	WHIM descriptors
RDF025u	radial distribution function-2.5/unweighted	RDF descriptors
QYYe	Qyy COMMA2 value/weighted by atomic Sanderson elecronegativities	geometrical descriptors
MATS6v	Moran autocorrelation lag 6/weighted by atomic van der Waals volumes	2D autocorrelation
RDF085m	radial distribution function-8.5/weighted by atomic masses	RDF descriptors
R3e+	R maximal autocorrelation of lag3/weighted by atomic Sanderson elecronegativities	GETAWAY descriptors
ISIZ	information index of molecular size	topological descriptors
G2e	2nd component symmetry directional WHIM index/weighted by atomic Sanderson elecronegativities	WHIM descriptors
RDF105u	radial distribution function-10.5/unweighted	RDF descriptors
MATS1p	Moran autocorrelation lag-1/weighted by atomic polarizabilities	2D autocorrelation
ZM1V	first Zagreb index by valence vertex degree	topological descriptors

Table 5. Dragon Descriptors Selected by Fuzzy Rough Set ACO

symbol	definition	class
dp01	molecular profile no.01	Randic molecular profile
Mor11u	3D-MoRSE-signal 11/unweighted	3D MoRSE descriptors
Mor07v	3D-MoRSE-signal 7/weighted by atomic van der Waals volumes	3D MoRSE descriptors
X1sol	solvation connectivity index chi-0	topological descriptors
GATS2v	Geary autocorrelation lag 2/weighted by atomic van der Waals volumes	2D autocorrelation
MPC06	molecular path count of order 06	topological descriptors
RDF025v	radial distribution function-2.5/weighted by atomic van der Waals volumes	RDF descriptors
BEHv	highest eigenvalue n.1 Burden matrix/weighted by atomic van der Waals volumes	BCUT descriptors
Mor06u	3D-MoRSE-signal 06/unweighted	3D MoRSE descriptors
IC5	information content index (neighborhood symmetry of −5 order)	topological descriptors
RTv+	R matrix index/weighted by atomic van der Waals volumes	GETAWAY descriptors
X1Av	average valence connectivity index chi-1	topological descriptors
GATS2p	Geary autocorrelation lag 2/weighted by atomic polarizabilities	2D autocorrelation
MATS6p	Moran autocorrelation lag-6/weighted by atomic polarizabilities	2D autocorrelation
Mor31p	3D-MoRSE-signal 31/weighted by atomic polarizabilities	3D MoRSE descriptors
PJI3	3D Petijean shape index	geometrical descriptors

above procedure, a Y-scrambling test was performed, where the output values of the compounds were shuffled randomly, and the scrambled data set was re-examined by the MLR method against real (unscrambled) input descriptors to determine the correlation and predictability of the resulting 'model'. The results of Y-scrambling are shown in Table 1. Data treatment, namely variable selection and regression, was achieved by using the Matlab platform.[27] The predictive ability of the various methods tested were evaluated through the following statistical parameters: squared correlation coefficient of the experimental versus fitted or predicted bioactivity values (r^2 and r^2_{test}), root-mean-square errors of calibration and external validation (RMSEP), relative standard error of prediction (RSEP), mean absolute error (MAE), Fischer test (F), and t-test.

3. RESULTS AND DISCUSSION

A series of Dragon descriptors was obtained for each GSK-3β inhibitor of Table 2. Only sixteen descriptors from the total pool were independently selected by suitable selection methods, namely genetic algorithm (GA), successive projections algorithm (SPA), and fuzzy rough set ant colony optimization (fuzzy rough set ACO). Such descriptors are properly depicted in Tables 3 −5 and were regressed against the biological activities using multiple linear regression (MLR), artificial neural network (ANN), and support vector machines (SVM).

In the ANN generation, the data set was separated into three groups: training, validation set, and test sets. All molecules were placed in these sets based on their activities. The training set, consisting of 100 molecules (65% of whole data set), was used for the model generation. However, the validation set, consisting of 23 molecules (15% of whole data set), was used to take care on the overtraining. The test set, consisting of 29 molecules (20% of whole data set), was used to evaluate the generated model. A three-layer network with a sigmoid transfer function was designed for ANN. Before training the networks, the input and output values

Table 6. Statistical Parameters of the QSAR Modeling Using Genetic Algorithm (GA) As a Variable Selection Method

parameter		MLR[a]	ANN	SVM
r^2	training set	0.770	0.889	0.929
	validation set		0.875	
	test set	0.738	0.888	0.910
RMSEP	training set	0.413	0.281	0.232
	validation set		0.339	
	test set	0.482	0.324	0.274
RSEP (%)	training set	5.918	4.019	3.321
	validation set		4.867	
	test set	7.047	4.740	3.996
MAE (%)	training set	5.123	4.245	3.849
	validation set		10.908	
	test set	11.249	9.573	8.585
F	training set	406.066	787.438	1586.263
	validation set		147.593	
	test set	76.114	213.822	273.805
t-test	training set	20.151	28.061	39.828
	validation set		12.149	
	test set	8.724	14.623	16.547
σ				0.831
ϵ				0.07
C				33
no. of nodes in the input layer			16 + 1[b]	
no. of nodes in the hidden layer			7	
no. of nodes in output layer			1	
learning rate (μ)			0.65	
momentum			0.79	
transfer function			sigmoid	
no. of iterations (λ)			14	

[a] MLR equation: $pIC_{50} = -28.366 - 0.030832$ Mor01p + 132.25 GATS8m − 4.2165 MATS3p + 0.59764 Se − 1.5651 GGI3 − 20.651 JGI1 + 0.034925 TIC2 − 4.6951 MATS6v + 0.27881 Mor06e + 10.382 Ms −1 0.0282 qpmax − 36.986 $R5v^+$ − 0.11377 RDF025e + 29.294 $R6u^+$ + 0.30187 MWC02 − 0.048041 ZM1V. [b] Bias.

Table 7. Statistical Parameters of the QSAR Modeling Using Successive Projections Algorithm (SPA) As a Variable Selection Method

parameter		MLR[a]	ANN	SVM
r^2	training set	0.773	0.883	0.926
	validation set		0.878	
	test set	0.671	0.852	0.912
RMSEP	training set	0.411	0.293	0.247
	validation set		0.339	
	test set	0.476	0.322	0.281
RSEP (%)	training set	5.885	4.190	3.545
	validation set		4.869	
	test set	6.954	4.710	4.105
MAE (%)	training set	5.105	4.752	4.067
	validation set		10.614	
	test set	11.430	9.843	8.844
F	training set	411.863	738.901	1521.222
	validation set		151.499	
	test set	54.999	155.684	277.961
t-test	training set	20.294	27.183	39.003
	validation set		12.309	
	test set	7.416	12.477	16.672
σ				0.432
E				0.06
C				17
no. of nodes in the input layer			16 + 1[b]	
no. of nodes in the hidden layer			7	
no. of nodes in output layer			1	
learning rate (μ)			0.775	
momentum			0.678	
transfer function			sigmoid	
no. of iterations (λ)			10	

[a] MLR equation: $pIC_{50} = -30.625 - 1.0743$ Mor20m + 5.6787 P1p + 1.8036 Sv + 2.6699 Se − 1.6642 Sp + 24.075 G2u − 0.1055 RDF025u + 0.013861 QYYe − 3.6404 MATS6v − 0.089737 RDF085m + 11.256 R3e −0.3422 ISIZ + 24.941 G2e − 0.11185 RDF105u + 4.9133 MATS1p − 0.031276 ZM1V. [b] Bias.

were normalized between −1 and 1. The initial weights were selected randomly between −0.3 and 0.3. The network was then trained using the training set by the back-propagation strategy for optimization of the weights and bias values. The proper number of nodes (neurons) in the hidden layer was determined by training the network with a different number of nodes in the hidden layer. The root-mean-square error (RMSE) value measures how good the outputs are in com-

Table 8. Statistical Parameters of the QSAR Modeling Using Fuzzy Rough Set ACO As a Variable Selection Method

parameter		MLR[a]	ANN	SVM
r^2	training set	0.806	0.932	0.960
	validation set		0.928	
	test set	0.760	0.915	0.927
RMSEP	training set	0.379	0.229	0.179
	validation set		0.279	
	test set	0.424	0.289	0.240
RSEP (%)	training set	5.435	3.280	2.571
	validation set		4.004	
	test set	6.201	4.228	3.510
MAE (%)	training set	4.833	3.882	3.402
	validation set		9.674	
	test set	10.651	9.152	7.827
F	training set	503.812	1342.635	2889.918
	validation set		271.614	
	test set	85.398	288.873	342.421
t-test	training set	22.446	36.642	53.758
	validation set		16.481	
	test set	9.241	16.996	18.505
σ				0.25
E				0.04
C				8
no. of nodes in the input layer			16 + 1[b]	
no. of nodes in the hidden layer			7	
no. of nodes in output layer			1	
learning rate (μ)			0.37	
momentum			0.49	
transfer function			sigmoid	
no. of iterations (λ)			17	

[a] MLR equation: $pIC_{50} = -33.599 - 37.787$ DP10 $- 0.15212$ Mor11u $+ 0.39573$ Mor07v $+ 89.358$ X1sol $+ 1.2498$ GATS2v $- 0.11021$ MPC06 $+ 0.012769$ RDF025v $+ 2.241$ BEHv8 $- 0.165$ Mor06u $+ 0.038866$ IC5 $+ 7.1944$ $RTv^+ + 2.3517$ X1Av $+ 4.9369$ GATS2p $-$ 2.7313 MATS6p $+ 0.28866$ Mor31p $- 5.2234$ PJI3. [b] Bias.

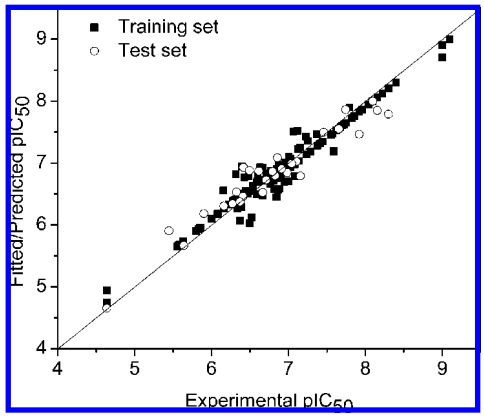


Figure 3. Plot of experimental vs fitted/predicted pIC_{50} for the series of GSK-3β inhibitors, obtained through the QSAR model based on the fuzzy rough set ACO selection method and SVM regression.

Table 9. Statistical Parameters of the QSAR Modeling Using PLS As a Regression Method

parameter		SPA-PLS	GA-PLS	fuzzy rough set ACO-PLS
r^2	training set	0.781	0.783	0.817
	test set	0.696	0.747	0.803
q^2		0.758	0.773	0.794
PLS components		8	8	7
RMSEP	training set	0.406	0.421	0.379
	test set	0.472	0.445	0.369
RSEP (%)	training set	5.823	6.031	5.426
	test set	6.890	6.495	5.386
MAE (%)	training set	4.964	5.215	4.980
	test set	11.583	11.179	10.259
F	training set	432.480	436.390	539.577
	test set	61.925	79.679	110.113
t-test	training set	20.796	20.890	23.229
	test set	7.869	8.926	10.493

parison with the target values. It should be noted that for evaluating the overtraining, the training of the network for the prediction of pIC_{50} must stop when the RMSE of the validation set begins to increase while RMSE of training set continues to decrease. Therefore, training of the network was stopped when overtraining began. All of the above-mentioned steps were carried out using back-propagation, Levenberg−Marquardt update function. To select the best weight update function, two statistical methods were considered for evaluating the models developed using these function, namely leave-one-out cross-validation (evaluated by q^2) and the prediction standard error of estimation (SEP). Momentum from 0.3 to 0.7 was used to prevent local minima; obviously, if it is too small, it cannot avoid local minima effectively, whereas if it is too large, it may overshoot the minimum and cause system instability.

While calibration with MLR yielded QSAR models only reasonably predictable, with r^2 ranging from 0.77 to 0.81 and r^2_{test} of 0.67 to 0.76, ANN and specially SVM were capable of estimating and predicting biological activities very accurately (Tables 6−8). These results demonstrate the nonlinear nature of the correlation between the GSK-3β inhibitory activities and the selected descriptors, which is not accounted for by simple multiple linear regressions. In SVM, a linear estimation is done in kernel-induced feature space ($y = w^T\phi(x) + b$, where w corresponds to the weights and ϕ denotes a feature map). In applications involving nonlinear regression, it is enough to change the inner product

$\langle\phi(x_i), \phi(x_j)\rangle$ of eq 1 by a kernel function and the ijth element of kernel matrix $\mathbf{K}$ equals $\mathbf{K}_{ij} = \phi(x_i)^{\mathrm{T}}\phi(x_j)$

$$y_j = \sum_{i=1}^{N} \alpha_i \varphi(x_i)^T \varphi(x_j) + b = \sum_{i=1}^{N} \alpha_i \langle \varphi(x_i), \varphi(x_j) \rangle + b \quad (2)$$

If this kernel function meets Mercer's condition,[28] the kernel implicitly determines both a nonlinear mapping, $x \rightarrow \phi(x)$, and the corresponding inner product $\phi(x_i)^{\mathrm{T}}\phi(x_j)$. This leads to the following nonlinear regression function

$$y = \sum_{i=1}^{N} \alpha_i \mathrm{K}(x_i, x) + b \quad (3)$$

For a point x_j to be evaluated it is

$$y_j = \sum_{i=1}^{N} \alpha_i \mathrm{K}(x_i, x) + b \quad (4)$$

It is worth mentioning that the quality of prediction of SVMR is dependent on some kernel type parameters, which determine the sample distribution in the mapping space, and its corresponding parameters γ, controller of trade off C, and ε-insensitive loss function. The parameters of SVMR were optimized by systemically changing their values in the training step and calculating the RMSE and accuracy of the model using 5-fold cross-validation. The ε optimum value is supported by the type of noise that is usually unknown in the data, but there will be some feasible respectfulness of the number of resulting support vector if enough information of the noise is attainable to select an optimal value for ε. In fact, choosing the suitable value of ε is a critical step because ε-insensitivity averts meeting boundary conditions on training set and permits the possibility of sparsity in the dual formulations solution. On the other hand, the other parameter is regularization parameter C that controls the trade-off between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data; likewise if C is too large, then the SVM model will overfit on the training data.

This procedure gave excellent correlation both in calibration and prediction and thus may be reliably used to predict the bioactivities of novel, proposed compounds. Furthermore, when coupled to fuzzy rough set ACO (Table 8 and Figure 3), the statistical results were found to be superior to the remaining methods applied for the series of GSK-3β inhibitors, suggesting the application of this new feature selection method, which is based on fuzzy logic and group decision, in chemistry and other QSAR studies. In fact, this approach behaved more satisfactorily than the one in which the well-established partial least-squares (PLS) regression method, usually applied in 3D-QSARs, was applied to regress Dragon descriptors against the bioactivities of the GSK-3β inhibitors. According to PLS-based models (Table 9), r^2 varied from 0.78 to 0.82 for the training set and from 0.70 to 0.80 for the test set. Leave-one-out (LOO) cross-validation experiments gave q^2 ranging from 0.76 to 0.79. These results are only comparable to MLR-based models, independent of the feature selection method used.

Overall, support vector machines (SVM) are suggested to be used as a regression method in QSAR studies where linear behavior is not expected or in those studies in which linear approaches do not work well. In addition, fuzzy rough set ACO showed the potential of extracting relevant information from a variety of descriptors, given the encouraging results obtained in its first use in chemistry here.

ACKNOWLEDGMENT

CNPq is gratefully acknowledged for the fellowship (to M.P.F.).

REFERENCES AND NOTES

(1) Cramer, R. D., III.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959.
(2) Klebe, G.; Abraham, U.; Mietzner, T. *J. Med. Chem.* **1994**, *37*, 4130.
(3) Goodford, P. J. *GRID*; University of Oxford: Oxford, UK, 1995.
(4) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. *J. Am. Chem. Soc.* **1997**, *119*, 10509.
(5) Vedani, A.; Dobler, M. *J. Med. Chem.* **2002**, *45*, 2139.
(6) Vedani, A.; Dobler, M.; Lill, M. A. *J. Med. Chem.* **2005**, *48*, 3700.
(7) Freitas, M. P.; Martins, J. A.; Brown, S. D. *J. Mol. Struct.* **2005**, *738*, 149.
(8) Freitas, M. P. *Org. Biomol. Chem.* **2006**, *4*, 1154.
(9) Freitas, M. P. *Curr. Comput.-Aided Drug Des.* **2007**, *3*, 235.
(10) Tian, F.; Zhou, P.; Li, Z. *J. Mol. Struct. (Theochem)* **2007**, *871*, 140.
(11) *Handbook of Genetic Algorithms*; Davis, L., Ed.; Van Nostrand-Reinhold: London, 1991.
(12) Zhang, J.; Rivard, B.; Rogge, D. M. *Sensors* **2008**, *8*, 1321.
(13) Bonabeau, E.; Dorigo, M.; Theraulez, G. *Swarm Intelligence: From Natural to Artificial Systems*; Oxford University Press Inc.: New York, 1999.
(14) Jensen, R. In *Studies in Computational Intelligence*; Abraham, A., Grosan, C., Ramos, V., Eds.; Springer: Heidelberg, 2006; Vol. 34, pp 45−73.
(15) Jensen, R.; Shen, Q. *Fuzzy Sets Sys.* **2005**, *149*, 5.
(16) Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning about Data*; Kluwer Academic Publishing: Dordrecht, 1991.
(17) Polkowski, L. *Rough Sets: Mathematical Foundations. Advances in Soft Computing*; Physica Verlag: Heidelberg, Germany, 2002.
(18) Chouchoulas, A.; Shen, Q. *Appl. Art. Intell.* **2001**, *15*, 843.
(19) Jensen, R.; Shen, Q. *IEEE Trans. Know. Data Eng.* **2004**, *16*, 1457.
(20) Mager, P. P. *Med. Chem. Res.* **1998**, *8*, 277.
(21) Belousov, A. I.; Verzakov, S. A.; von Frese, J. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 15.
(22) Belousov, A. I.; Verzakov, S. A.; von Frese, J. *J. Chemom.* **2002**, *16*, 482.
(23) Thissen, U.; van Brakel, R.; de Weijer, A. P.; Melssen, W. J.; Buydens, L. M. C. *Chemom. Intell. Lab. Syst.* **2003**, *69*, 35.
(24) Taha, M. O.; Bustanji, Y.; Al-Ghussein, M. A. S.; Zaloum, H.; Al-Masri, I. M.; Atallah, N. *J. Med. Chem.* **2008**, *51*, 2062.
(25) *HyperChem version 7.0*, Hypercube, Inc.: Gainesville, 2007.
(26) Todeschini, R.; Consonni, V.; Pavan, M. *Dragon software*; 2002.
(27) *Matlab version 7.5*; MathWorks Inc.: Natick, 2007.
(28) Mercer, J. *Philos. Trans. R. Soc. London A* **1909**, *209*, 415.

CI9000103