Sweetness power QSARs by PRECLAV software

Tarko Laszlo,[a*] Irina Lupescu,[b] and Diana Groposila-Constantinescu [b]

[a] Center of Organic Chemistry "C.D.Nenitzescu" – Romanian Academy, Spl. Independentei 202B, 6th Sector, Bucharest, PO Box 15-258, MC 60023, Fax 3121601
[b] Center of Applied Biochemistry and Biotechnologies "BIOTECHNOL", B-dul Marasti 59, 011464,1st Sector, Bucharest, Fax 2242815
E-mail: ltarko@cco.ro

Dedicated to Professor Alexandru T. Balaban on his 75th birthday
(received 18 Mar 05; accepted 08 Jun 05; published on the web 30 Jun 05)

Abstract
This paper presents some QSAR (*Quantitative Structure Activity Relationship*) studies with a testing set, realized by the PRECLAV (*Property Evaluation by Class Variables*) computer program. The database we used contains sweeteners with very diverse structures – sugars, halosugars, guanidine derivatives and 3-aminosuccinamic acid derivatives. According to their estimated values of Log(RS), the testing set molecules are classified as "recommended", "uncertain", or "un-recommended" for synthesis. Comparing the estimated Log(RS) values with the observed values we have found that the aforementioned classification is sufficiently correct to have actual practical value, even if the training/testing set contains sweeteners of several different classes. The N-phenyl-guanidine-acetic acid derivatives, with a polycyclic system bonded with the nitrogen atom, represent a distinct subclass of guanidinic sweeteners.

Keywords: QSAR, PRECLAV, sweeteners

Introduction

The PRECLAV (*Property Evaluation by Class Variables*) computer program[32] has been used for several years in doing QSAR (*Quantitative Structure Activity Relationship*) studies for "academic" purposes (to test the quality of certain algorithms and/or the predicting ability of certain descriptors) as well as to solve "practical" problems that have been proposed by various research groups in the drug design area (identifying the predictors having the highest influence on the values of the dependent property, and estimating the value of the desired property for molecules not yet synthesized)[1-11].

We have recently thoroughly described the program's latest version algorithm.[12]

The present paper presents the results of some QSAR studies in which we have used databases containing sweeteners with a very diverse structure – sugars, halosugars, guanidine derivatives and dipeptides.

Methods and formulae

The molecules have been constructed virtually using the molecular mechanics program, PCMODEL[13].

The geometry of the minimum energy conformer was obtained by using the MMX force field and GMMX algorithm[14]. Further, the geometry was more rigorously optimized with the quantum mechanics program MOPAC[15], using the keyword string: "am1 pulay gnorm=0.01 shift=50 geo-ok mmok camp-king bonds vectors".

The *output* files created by MOPAC for each analyzed molecule are *input* files for PRECLAV and they contain the values of some descriptors. Using the data from the files generated by MOPAC, PRECLAV has computed most of the descriptors and has performed the statistical analysis. A detailed list of descriptors is available as supplementary material.

The analyzed dependent property was Log(RS), where RS (*relative sweetness*) is the sweetness power relative to sucrose. When the analyzed molecules had a common skeleton we used "whole molecule" and "grid" descriptors. Otherwise we used only "whole molecule" descriptors.

The QSAR studies can be made with or without a testing set. In the case of QSAR studies with a testing set, PRECLAV uses the Class function for identifying the significant descriptors. The QSAR equation that PRECLAV uses for prediction purposes in such situations is not the same as the equation one obtains when the program works without a testing set.

The "significant" descriptors satisfy conditions (**1**) and (**2**):

$$C_v > 3 \quad (1)$$

$$Q > 1 \quad (2)$$

where C_v is the coefficient of variation for descriptor values, defined as usual by

$$C_v = 100 \, x \, \sigma / V_m \quad (3)$$

where σ is the standard deviation around the average value, V_m is the average absolute value, and Q is the quality function for the analysed descriptor

$$Q = r^2 / [1 - C^a (1 - b \, x \, r^2_{min})] \quad (4)$$

where r^2 is the square of the Pearson linear correlation between the descriptor values and the dependent property values, r^2_{min} is the minimum value imposed for r^2; the default value for r^2_{min},

empirically established, is 4 / N (where N is the number of molecules from the training set); the user may modify this value, and C is Class function

$$C = \sigma_N / \sigma_{N+K} \; if \; \sigma_N < \sigma_{N+K} \quad (5a)$$

$$C = \sigma_{N+K} / \sigma_N \; if \; \sigma_N >= \sigma_{N+K} \quad (5b)$$

where σ_N is σ from formula (**3**) computed for N molecules from the training set, σ_{N+K} is σ from formula (**3**) computed for the entire database (N molecules from the training set + K molecules from the testing set), *a* is a real number, whose value is established empirically (*a* = 10) by analysing a large number of databases (training set + testing set), and *b* =1 for the "whole molecule" descriptors and *b* = 2 for the "grid" descriptors (this way the "grid" descriptors selection is more drastic)

It is considered that the Class function measures how representative a sample – from the statistical point of view - is the training set in the joint set of the testing and training sets from the analyzed descriptor's point of view. If the testing set is missing then C = 1 for all descriptors and the condition (**4**) becomes $r^2 > b \times r^2_{min}$.

Usually, according to the selection criteria (**1**) and (**2**), only 5–25% of the computed descriptors are "significant".

The results of some QSAR studies performed without a testing set, using the same databases we have used here, will be presented in a future paper. Here we present only the results of several QSAR studied performed with a testing set. The training and testing sets have been defined by a standard procedure. This procedure involves the ordering of the molecules in the database according to the value of the dependent property, starting with the smallest value. The molecules with rank 3, 8, 13, 18, 23 … in the string will form the actual testing set.

The analysis of the training set molecules has produced tens of thousands of multi-linear QSAR equations of the following form:

$$Log(RS) = c_0 + \sum c_k \cdot p_k \quad (6)$$

The "best" QSAR equation was selected according to the value of a cross-validation quality function, specific to PRECLAV[12]. This equation was then utilized for predicting the values of Log(RS) for the molecules in the testing set. Once the computations were over, the testing set molecules were classified in three categories: "recommended for synthesis", "uncertain", and "un-recommended for synthesis". The classification was based on Log(RS)'s estimated value, relative to the other estimated values for the rest of the molecules in the testing set. After computing the values of the dependent property for the molecules in the testing set, PRECLAV sorts these molecules according to the estimated values. An average value P_{calc}^m is computed for the estimated values and also a standard deviation σ of the estimated values around the average.

The program considers "high" the value fulfilling the criterion (**7**) and "low" the value fulfilling the criterion (**8**):

$$P_{calc} > P_{calc}^{m} + 0.5 \times \sigma \quad (7)$$

$$P_{calc} < P_{calc}^{m} - 0.5 \times \sigma \quad (8)$$

If the user wishes to synthesize molecules with a pronounced biochemical activity, the molecules fulfilling criterion (**7**) are "recommended for synthesis", while the ones fulfilling criterion (**8**) are "un-recommended for synthesis".

In the "practical" QSAR studies, the testing set contains new molecules, not yet synthesized, with a structure imagined by the program user. In this case the observed values of Log(RS) for the testing set are not known because the molecules have not yet been analyzed by physical / chemical methods. It is very important that the program properly sorts the testing set molecules by the estimated values of Log(RS), even if the values themselves do not correspond too well with the real values – the most important thing is that the program arranges the molecules in the correct order. This way the molecules "recommended for synthesis" can be correctly identified. Thus, in the "academic" QSAR studies we present here, we have considered that an adequate measure for the quality of the prediction is the value of the Kendall rank correlation between the computed and the observed values of Log(RS).

The SMILES notation of analysed molecules is available as supplementary material.

Results and Discussion

QSAR study #1

Database: sugars and halosugars, 41 molecules (Fig. 1, Table 1)
Dependent property: Log(RS), the values are taken from literature[16, 17]
Training set: 33 molecules (Table 1, normal font)
Testing set: 8 molecules (Table 1, bold font)
Descriptors: "whole molecule" + "grid"
Number of significant descriptors: 99

*The type (**6**) QSAR equation for prediction:*

$c_0 = .0019$
$c_1 = .7241$
p_1 – QSAR of molecular orbital energies
$c_2 = -1.8699$
p_2 – A121 (electrostatic attraction force, "grid" descriptor)
$c_3 = -104.4192$
p_3 – F100 (electrostatic resultant force, "grid" descriptor)

Standard error (training set): 0.338
Fisher F function (training set): 139.9
Kendall cross-validated rank correlation K_{CV} (training set) = 0.8788
Kendall rank correlation K (testing set): 0.7143
Standard error (testing set): 0.489

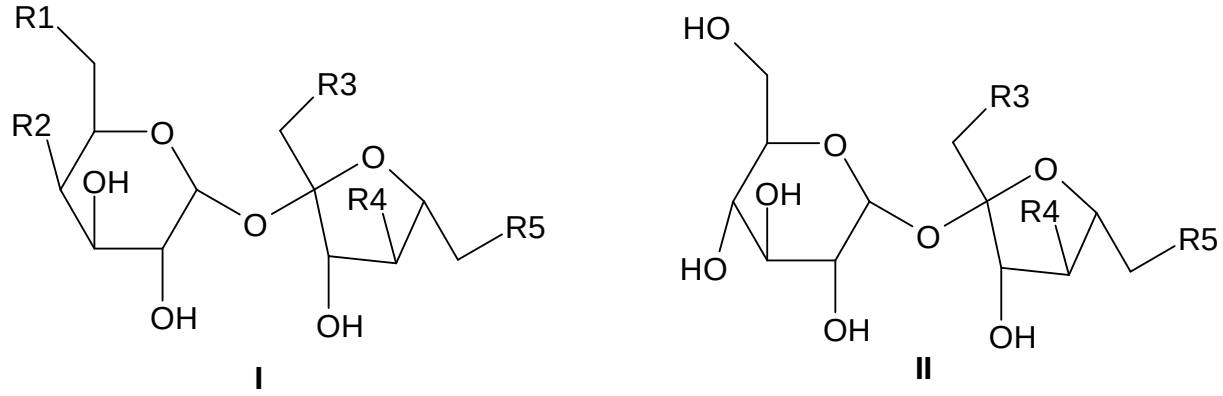

Figure 1. Structure of sugars/halosugars.

The three molecules in the testing set having the smallest observed values of Log(RS) have been labeled "un-recommended for synthesis". Two molecules having the highest observed values of Log(RS) have been labeled "recommended for synthesis" and another one has been labeled "uncertain". In case of molecule **23** the value of Log(RS) is over-estimated, while for molecule **33** the value of Log(RS) is under-estimated.

Table 1. Log(RS) values of sugars/halosugars

Crt. No.	Name	Str. in Fig. 1	R_1	R_2	R_3	R_4	R_5	Log RS Obs.	Log RS Calc.	Recommended for synthesis
1	Lactose	-	-	-	-	-	-	-0.699	-0.151	
2	-	I	OH	OH	OH	OH	OH	-0.699	-0.174	
3	**Mannose**	**-**	**-**	**-**	**-**	**-**	**-**	**-0.523**	**-0.081**	**no**
4	Galactose	-	-	-	-	-	-	-0.495	-0.819	
5	Maltose	-	-	-	-	-	-	-0.481	0.250	
6	Xylose	-	-	-	-	-	-	-0.398	-0.204	
7	α-Glucose	-	-	-	-	-	-	-0.155	-0.478	
8	**β-Glucose**	**-**	**-**	**-**	**-**	**-**	**-**	**-0.097**	**-0.523**	**no**
9	Sorbose	-	-	-	-	-	-	-0.066	0.242	
10	-	I	OH	H	OH	OH	OH	0.000	0.121	
11	-	II	-	-	OH	OH	OH	0.000	-0.220	
12	Fructose	-	-	-	-	-	-	0.236	0.007	
13	**-**	**I**	**OH**	**OH**	**OH**	**Cl**	**OH**	**0.301**	**0.529**	**no**
14	-	I	OH	Cl	OH	OH	OH	0.699	0.884	
15	-	I	OH	OH	OH	Cl	Cl	0.699	1.087	
16	-	I	OH	OH	Cl	OH	OH	1.301	0.413	
17	-	I	OH	OH	OH	OH	Cl	1.301	1.191	
18	**-**	**II**	**-**	**-**	**Cl**	**OH**	**OH**	**1.301**	**0.923**	**uncertain**
19	-	II	-	-	OH	OH	Cl	1.301	1.250	
20	-	I	Cl	OH	Cl	OH	Cl	1.398	1.226	
21	-	I	OH	OH	Cl	Cl	OH	1.477	1.287	
22	-	I	OH	F	F	OH	F	1.602	1.579	
23	**-**	**I**	**OH**	**Cl**	**OH**	**OH**	**Cl**	**1.699**	**2.073**	**yes**
24	-	I	OH	OH	Cl	OH	Cl	1.881	1.608	
25	-	II	-	-	Cl	OH	Cl	1.903	1.501	
26	-	II	-	-	Br	OH	Br	1.903	1.910	
27	-	I	OH	OH	Cl	Cl	Cl	2.000	2.173	
28	**-**	**II**	**-**	**-**	**Cl**	**Cl**	**Cl**	**2.000**	**2.050**	**yes**
29	-	I	OH	Cl	Cl	OH	OH	2.079	1.717	
30	-	I	OH	Cl	Cl	H	Cl	2.176	1.950	
31	-	I	OH	Cl	OH	Cl	Cl	2.204	2.796	
32	-	I	Cl	Cl	Cl	OH	Cl	2.301	2.338	
33	**-**	**I**	**OH**	**Cl**	**Cl**	**Cl**	**OH**	**2.342**	**1.498**	**uncertain**
34	-	I	OH	Br	Cl	OH	Cl	2.574	2.836	
35	-	I	H	Cl	Cl	OH	Cl	2.602	2.673	
36	-	I	OH	Cl	Cl	OH	Cl	2.813	2.882	
37	-	I	OH	Cl	Br	OH	Br	2.903	2.918	
38	**-**	**I**	**OH**	**Cl**	**Cl**	**F**	**Cl**	**3.000**	**2.900**	**yes**
39	-	I	OH	Cl	Cl	Cl	Cl	3.477	3.062	
40	-	I	OH	Cl	Cl	Br	Cl	3.477	3.209	
41	-	I	OH	Cl	Cl	I	Cl	3.875	3.991	

In QSAR study # 1 the descriptor having the highest influence on the Log(RS) value is the "QSAR of molecular orbital energies". This descriptors gives Log(RS) as a linear function of the inverse of the energy differences between the HOMO-1, HOMO, LUMO and LUMO+1 molecular orbitals. When all the molecules from Table 1 had been included in the training set, the same descriptor proved to have the highest influence on the Log(RS) value. This suggests that the Log(RS) value for (halo)sugars correlates with the absorbed radiation wavelengths in the UV-VIS domain.

QSAR study #2

Database: guanidine derivatives, 41 molecules (Fig. 2, Table 2)
Dependent property: Log(RS), the values are taken from literature[16]
Training set: 33 molecules (Table 2, normal font)
Testing set: 8 molecules (Table 2, bold font)
Descriptors: "whole molecule" + "grid"
Number of significant descriptors: 21

*The type (**6**) QSAR equation for prediction:*

c_0 = .7139
c_1 = -32.6107
p_1 – A33 (electrostatic attraction force, "grid" descriptor)
c_2 = -1749.6464
p_2 – F120 (electrostatic resultant force, "grid" descriptor)
c_3 = -15.4273
p_3 – A64 (electrostatic attraction force, "grid" descriptor)

Standard error (training set): 0.356
Fisher F function (training set): 24.4
Kendall cross-validated rank correlation K_{CV} (training set) = .5758
Kendall correlation K (testing set): 0.7857
Standard error (testing set): 0.393

The three testing set molecules having the highest values of Log(RS) have been labeled "recommended for synthesis". The molecule having the smallest Log(RS) value has been labeled "un-recommended for synthesis".

It is remarkable how few significant descriptors there are. Due to how PRECLAV selects the significant descriptors (from a group of almost 1000 computed), a small number of significant descriptors suggest that the training set is not a representative sample for the molecules in Table 2. From the group of 21 retained significant descriptors only 5 are "grid" descriptors. Nevertheless, the equation utilized for prediction contains only "grid" predictors.

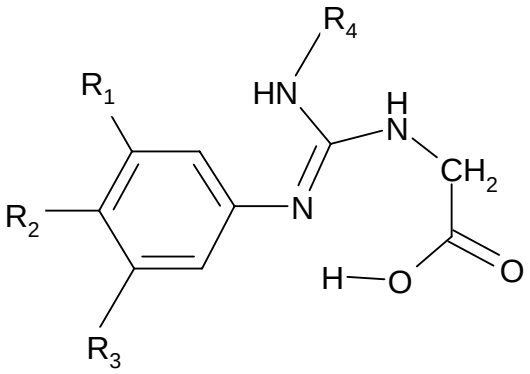

Figure 2. Structure of guanidine derivatives.

Table 2. Log(RS) values of guanidine derivatives

Crt. No.	R_1	R_2	R_3	R_4	LogRS Obs.	LogRS Calc.	Recommended for synthesis
42	H	CN	H	CH_2CH_3	2.544	2.891	
43	Cl	H	Cl	C_6H_3(3,5-diCl)	3.000	3.140	
44	**H**	**CN**	**H**	**H**	**3.431**	**2.762**	**no**
45	H	CN	H	C_6H_5	3.603	3.739	
46	H	CN	H	C_6H_4(2-CH_3)	3.699	3.886	
47	H	H	H	($CH(CH_3)C_6H_5$ *S*	3.699	4.404	
48	CN	H	H	($CH(CH_3)C_6H_5$ *S*	3.740	4.007	
49	**H**	**CN**	**H**	**$(CH_2)_5CH_3$**	**3.778**	**4.038**	**uncertain**
50	NO_2	H	H	($CH(CH_3)C_6H_5$ *S*	3.778	3.737	
51	H	CN	H	C_6H_4(4-CH_3)	3.845	3.991	
52	H	NO_2	H	($CH(CH_3)C_6H_5$ *S*	3.845	4.448	
53	CF_3	H	H	($CH(CH_3)C_6H_5$ *S*	3.875	4.085	
54	**H**	**CN**	**H**	**$(CH_2)_2C_6H_5$**	**3.929**	**4.261**	**uncertain**
55	H	CN	H	($CH(CH_3)C_6H_5$ *R*	3.954	4.408	
56	H	CN	H	C_6H_4(3-CH_3)	3.954	4.158	
57	H	CN	H	C_6H_4(3-Cl)	4.000	3.442	
58	H	CN	H	Cyc- C_6H_{11}	4.079	3.939	
59	**CH_3**	**H**	**H**	**($CH(CH_3)C_6H_5$ *S***	**4.079**	**4.504**	**uncertain**
60	F	H	F	($CH(CH_3)C_6H_5$ *S*	4.176	4.522	
61	Cl	H	Cl	cyc-C_7H_{13}	4.301	4.487	
62	Br	H	H	($CH(CH_3)C_6H_5$ *S*	4.398	4.268	
63	H	CN	H	($CH(CH_3)C_6H_5$ *S*	4.447	4.215	
64	**H**	**CN**	**H**	**$CH_2C_6H_5$**	**4.477**	**4.190**	**uncertain**
65	CH_3	H	CH_3	($CH(CH_3)C_6H_5$ *S*	4.477	4.493	
66	H	CN	H	CH_2-(cyc-C_6H_{11})	4.544	4.689	
67	Cl	Cl	Cl	($CH(CH_3)C_6H_5$ *S*	4.544	4.346	
68	Cl	H	Cl	CH_2-(cyc-C_6H_{11})	4.544	4.805	
69	**H**	**CN**	**H**	**($CH(CH_3)$-cyc-C_6H_{11} *S***	**4.699**	**4.877**	**yes**
70	CH_3	CN	H	($CH(CH_3)C_6H_5$ *S*	4.699	4.309	
71	CH_3	CN	CH_3	($CH(CH_3)C_6H_5$ *S*	4.699	4.326	
72	H	CN	H	cyc-C_7H_{13}	4.778	4.502	
73	Cl	H	Cl	cyc-C_8H_{15}	4.778	4.591	
74	**Cl**	**H**	**Cl**	**($CH(CH_3)$-cyc-C_6H_{11} *S***	**4.845**	**4.820**	**yes**
75	Cl	H	Cl	$CH_2C_6H_5$	4.903	4.358	
76	Cl	H	Cl	($CH(CH_3)C_6H_5$ *S*	5.079	4.401	
77	H	CN	H	cyc-$C_{10}H_{19}$	5.176	5.703	
78	H	CN	H	$CH(C_6H_5)_2$	5.176	5.138	
79	**Cl**	**H**	**Cl**	**$CH(C_6H_5)_2$**	**5.204**	**5.122**	**yes**
80	H	CN	H	cyc-C_8H_{15}	5.230	4.620	
81	H	CN	H	$CH_2C_6H_4$(3-CH_3)	5.301	5.047	
82	H	CN	H	cyc-C_9H_{17}	5.301	5.069	

In QSAR study # 2 the descriptor having the highest influence on the Log(RS) value is the "grid" descriptor A33. When all the molecules from Table 2 had been included in the training set, the "bond orders sum" descriptor proved to have the highest influence on Log(RS). This suggests that in case of Fig. 2 guanidine derivatives the values of Log(RS) depend on the molecular size and on the un-saturation degree of the chemical bonds. The importance of the size of the molecule is stressed too – using the "moment of inertia C" descriptor – by the QSAR study on guanidines, performed with a very different training set by Katrizky et al.[33]

There have been synthesized some guanidines where the R_4 chemical group (see Figure 2) contains a polycyclic system (naphthyl, indanyl, adamantyl, 1,3-benzodioxolil etc.)[16]. We have performed numerous other QSAR studies using PRECLAV (that are not included here) with a database including both the molecules from Table 2 and several guanidines with a polycyclic system. No matter how we grouped the molecules in the training and testing sets, the prediction power of the resulting equations was much weaker – for both the training set and the testing set molecules. Therefore, we are drawing the conclusion that the guanidines with a R_4 containing a polycyclic system and the guanidines from Table 2 belong to two different subclasses of guanidinic sweeteners.

QSAR study #3

Database: 3-aminosuccinamic acid derivatives, 41 molecules (Fig. 3, Table 3)
Dependent property: Log(RS), values taken from literature[18-31]
Training set: 33 molecule (Table 3, normal font)
Testing set: 8 molecule (Table 3, bold font)
Descriptors: "whole molecule" + "grid"
Number of significant descriptors: 388

*The type (**6**) QSAR equation for prediction:*

c_0 = -20.8851
c_1 = .8607
p_1 – QSAR of molecular orbital energies
c_2 = -14.8874
p_2 – A34 (electrostatic attraction force, "grid" descriptor)
c_3 = 1.8796
p_3 – Platt topologic index / Heavy atoms number ratio
c_4 = 1.0863
p_4 – E(lumo+1) - E(homo-1) gap
c_5 = 6.6894
p_5 – R84 (electrostatic repulsion force, "grid" descriptor)

Standard error (training set): 0.193
Fisher F function (training set): 120.3
Kendall cross-validated rank correlation K_{CV} (training set) = .9129
Kendall correlation K (testing set): 0.5714
Standard error (testing set): 0.715

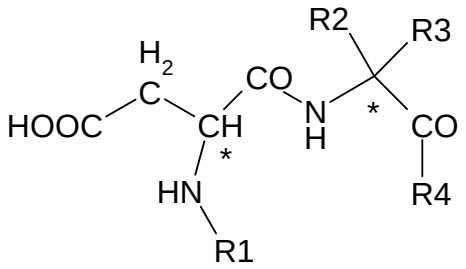

Figure 3. Structure of 3-aminosuccinamic acid derivatives.

Table 3. Log(RS) values of 3-aminosuccinamic acid derivatives

Crt. No.	R_1	R_2	R_3	R_4	Cfg. in Fig. 3	LogRS Obs.	Log RS Calc.	Reco-mmended for synthesis
83	H	H	CH_3	$O(CH_2)_3CH_3$	L – D	1.041	1.552	
84	H	H	CH_3	OCH_3	L – D	1.415	1.507	
85	**H**	**H**	**$(CH_2)_3CH_3$**	**OCH_3**	**L – D**	**1.613**	**2.117**	**no**
86	H	H	$CH(CH_3)(CH_2)_3CH_3$	OCH_3	L – D	1.785	1.771	
87	H	H	CH_3	OC_2H_5	L – D	1.908	1.674	
88	H	H	$(CH_2)_2C_6H_5$	OCH_3	L – L	2.004	2.234	
89	H	H	$CH(CH_3)_2$	NHCH(cyc-propyl)$_2$	L – D	2.045	2.240	
90	**H**	**CH_3**	**C_6H_5**	**OCH_3**	**L – L**	**2.179**	**2.890**	**uncertain**
91	H	H	CH_3	$O(CH_2)_2CH_3$	L – D	2.233	1.692	
92	H	H	CH_3	$NHCH(CH_3)(C_6H_5)$ *S*	L – D	2.258	2.265	
93	H	H	CH_3	O-cyc-hexyl	L – D	2.303	2.315	
94	H	CH_3	CH_3	$NHCH(C_2H_5)(C_6H_5)$ *S*	L – D	2.303	2.557	
95	**H**	**H**	**CH_2- cyclohexyl**	**OCH_3**	**L – D**	**2.354**	**2.833**	**uncertain**
96	H	H	cyc-hexyl	$NHCH(CH_2OCH3)(C_6H_5)$ *R*	L – D	2.400	2.396	
97	H	H	CH_3	$NHCH(CH_2OCH3)(C_6H_5)$ *R*	L – D	2.479	2.474	
98	H	H	CH_3	NH(2,6-diCH_3-C_6H_3)	L – D	2.700	2.647	
99	H	H	$CH(CH_3)(C_2H_5)$	$NHCH(C_2H_5)(C_6H_5)$ *S*	L – D	2.700	2.739	
100	**H**	**H**	**$CH(CH_3)_2$**	**$NHCH(CH3)(C_6H_5)$ *S***	**L – D**	**2.733**	**2.820**	**uncertain**
101	H	H	$COOCH_3$	O- cyc-pentyl	L – L	2.779	2.678	
102	H	H	$COOC_2H_5$	O-(2-CH_3- cyc-hexyl)	L – L	2.814	2.788	
103	H	H	CH_2-(bicyclo[2.2.1]-heptyl)	OCH_3	L – L	2.904	2.898	
104	$(CH_2)_3C_6H_5$	H	$CH_2C_6H_5$	OCH_3	L – L	3.000	3.157	
105	**H**	**H**	**2-furanyl**	**$NHCH(C_2H_5)(C_6H_5)$ *S***	**L – D**	**3.080**	**2.857**	**uncertain**
106	H	H	$CH(CH_3)_2$	$NHCH(C_2H_5)(C_6H_5)$ *S*	L – D	3.176	3.126	
107	H	H	C_2H_5	$NHCH(C_3H_7)(C_6H_5)$ *S*	L – D	3.301	2.981	

Table 3. Continued

108	H	H	C_2H_5	NHCH(CH_2OCH3) (C_6H_5) *R*	L – D	3.398	3.285	
109	$(CH_2)_2$-t-Bu	H	CH_3	NHCH(cyc-propyl)$_2$	L – D	3.398	3.567	
110	**$(CH_2)_2$-t-Bu**	**H**	**$CH(CH_3)_2$**	**NHCH(C_2H_5) (C_6H_5) *S***	**L – D**	**3.477**	**4.471**	**yes**
111	$(CH_2)_2$ $C(CH_3)_2$ C_6H_5	H	$CH_2C_6H_5$	OCH_3	L – L	3.602	3.412	
112	$(CH_2)_2$-t-Bu	H	C_2H_5	NHCH(CH_2OCH3) (C_6H_5) *R*	L – D	3.602	3.602	
113	$(CH_2)_2$-t-Bu	H	$CH(CH_3)_2$	NHCH(CH_2OCH3) (C_6H_5) *R*	L – D	3.602	3.487	
114	$(CH_2)_2$-t-Bu	CH_3	$CH_2C_6H_5$	OCH_3	L – L	3.740	3.655	
115	**H**	**H**	**$COOCH_3$**	**O-(2-CH_3-cyc-hexyl)**	**L – L**	**3.845**	**2.898**	**uncertain**
116	$(CH_2)_2$-t-Bu	H	C_2H_5	NHCH(C_2H_5)(C_6H_5) *S*	L – D	3.903	4.002	
117	$(CH_2)_3$-2,4-diOH-C_6H_3	H	$CH_2C_6H_5$	OCH_3	L – L	4.000	4.124	
118	$(CH_2)_3$-2,3,4-triOH-C_6H_2	H	$CH_2C_6H_5$	OCH_3	L – L	4.000	4.114	
119	H	H	CH_2-(2-furanyl)	OCH_3	L – D	4.000	4.045	
120	**$(CH_2)_3$-3,4-diOH-C_6H_3**	**H**	**$CH_2C_6H_5$**	**OCH_3**	**L – L**	**4.176**	**3.807**	**yes**
121	$(CH_2)_3$-3-OH,4-OCH_3-C_6H_3	H	$CH_2C_6H_5$	OCH_3	L – L	4.398	4.073	
122	$(CH_2)_3$-3,4,5-triOH-C_6H_2	H	$CH_2C_6H_5$	OCH_3	L – L	4.398	4.402	
123	$(CH_2)_2$-$C(CH_3)_2$- 3-OH,4-OCH_3-C_6H_3	H	$CH_2C_6H_5$	OCH_3	L – L	4.699	4.868	

The prediction for the training set molecules is very good (F and K_{CV} have high values).

The prediction for the testing set molecules is poorer (K = 0.5714). Nevertheless, molecule **85**, having the lowest Log(RS) value, is correctly labeled "un-recommended for synthesis", and molecule **120**, having the highest Log(RS) value, is correctly labeled as "recommended for synthesis".

In QSAR study # 3 the descriptor having the highest influence on the value of Log(RS) is the "E(lumo+1) - E(homo-1) gap" descriptor. When all the molecules from Table 3 were included in the training set, the "Platt topologic index / Heavy atoms number ratio" descriptor proved to have the highest influence on Log(RS). This suggests that in the case of dipeptides, the value of Log(RS) depends on the molecular size and on the ramification degree of catena.

QSAR study # 4

Database: 123 molecule (Table 1 + Table 2 + Table 3)

Dependent property: Log(RS)
Training set: 98 molecules (Table 1 + Table 2 + Table 3, without the testing set molecules)
Testing set: 25 molecules (Table 4)
Descriptors: "whole molecule"
Number of significant descriptors: 144

*The type (**6**) QSAR equation for prediction:*

c_0 = -.3004
c_1 = -.4185
p_1 – Number of O-H single or faint bonds
c_2 = -.1766
p_2 – Dipole moment (X component)
c_3 = .2548
p_3 – 100 * Max. atomic nucleophilic reaction index for C atoms
c_4 = 18.0493
p_4 – Number of triple bonds / Number of bonds ratio
c_5 = .0104
p_5 – Molecular weight

Standard error (training set): 0.485
Fisher F function (training set): 172.5
Kendall cross-validated rank correlation K_{CV} (training set) = 0.7921
Kendall correlation K (testing set): 0.7933
Standard error (testing set): 0.507

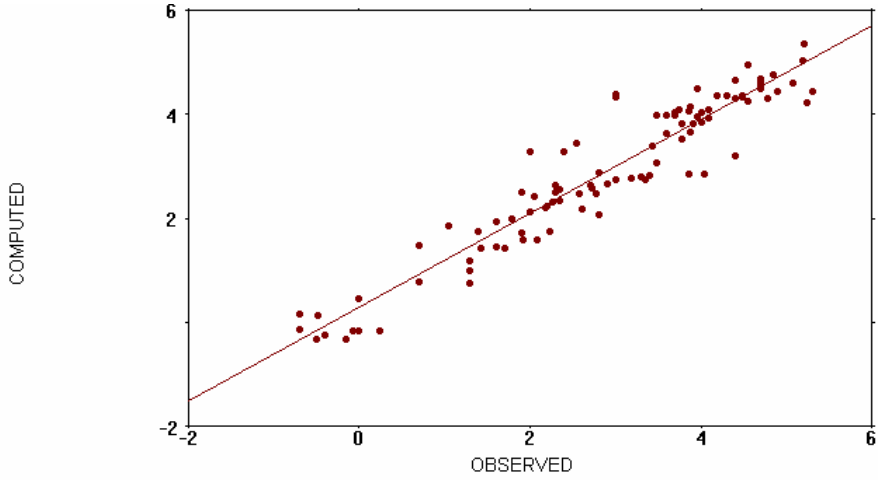

Figure 4. Observed/Computed values of Log(RS) - training set (entire database).

The "recommended for synthesis" group in testing set includes 8 guanidines and 2 dipeptides. The "un-recommended for synthesis" group in testing set includes 6 (halo)sugars and 1 dipeptide.

Table 4. Log(RS) values of testing set molecules (entire database)

Molecule	Obs. LogRS	Calc. LogRS	Calc. – Obs. difference	Recommended for synthesis
82	5.301	4.574	-0.727	yes
77	5.176	4.871	-0.305	yes
72	4.778	4.521	-0.257	yes
68	4.544	4.541	-0.003	yes
63	4.447	4.442	-0.005	yes
120	4.176	3.763	-0.413	yes
118	4.000	3.642	-0.358	uncertain
54	3.929	4.113	0.184	yes
52	3.845	4.076	0.231	yes
48	3.740	4.240	0.500	yes
111	3.602	4.542	0.940	yes
109	3.398	3.237	-0.161	uncertain
105	3.080	3.259	0.179	uncertain
37	2.903	2.664	-0.239	uncertain
99	2.700	2.962	0.262	uncertain
97	2.479	2.833	0.354	uncertain
93	2.303	2.162	-0.141	no
28	2.000	2.381	0.381	uncertain
24	1.881	1.357	-0.524	no
85	1.613	2.986	1.373	uncertain
21	1.477	1.905	0.428	no
17	1.301	0.618	-0.683	no
13	0.301	0.822	0.521	no
8	-0.097	-0.236	-0.139	no
3	-0.523	-0.179	0.344	no

In QSAR study # 4 the descriptor having the highest influence on the Log(RS) value is the "Molecular weight" descriptor. When all the molecules from Table 1, Table 2, and Table 3 were included in the training set, the "Percent of oxygen * Maximum charge of oxygen atoms product" descriptor proved to have the highest influence on Log(RS). This suggests that the value of Log(RS) depends on the size of molecules and on the electrostatic interactions involving oxygen atoms.

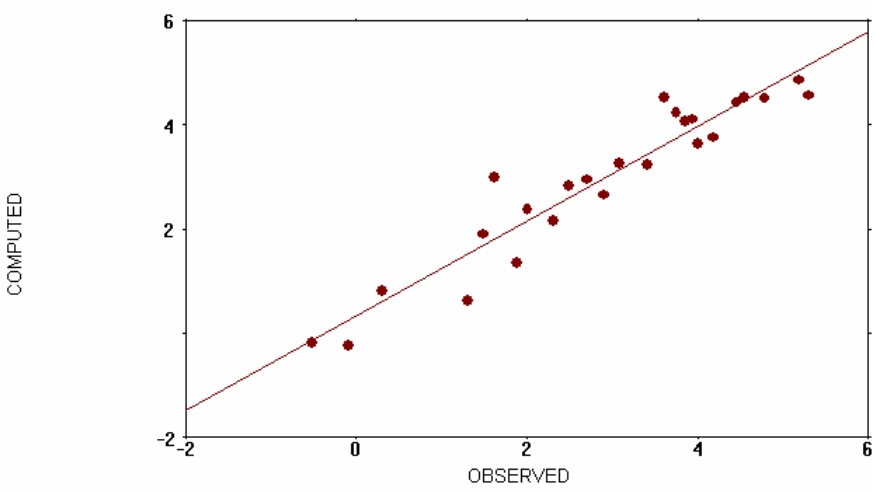

Figure 5. Observed/Computed values of Log(RS) - testing set (entire database).

Conclusions

PRECLAV software classifies the potential sweeteners from the testing set according to the values of Log(RS), in "recommended" or "un-recommended" for synthesis. By comparing the estimated values with the observed Log(RS) values we have found that the classification is mostly correct and thus it has practical value. This is the case even if the training/testing set contains sweeteners from several different classes.

The descriptors having the highest influence on Log(RS) are specific to each class of sweeteners.

The N-phenyl-guanidine-acetic acid derivatives, with a polycyclic system bonded to the nitrogen atom, represent a distinct subclass of N-phenyl-guanidine-acetic acid derivative sweeteners.

Supplementary Material is Available

Global descriptors
Grid descriptors

References

1. Tarko, L. *Rev.Chim.(Bucuresti)* **1997**, *48*, 676.
2. Crangus, C.; Nicolae, L.; Tarko, L. Sinteze de noi pesticide ditiofosforice concepute pe baza ecuatiilor QSAR, PROPLANT – Calimanesti-Romania 1998.
3. Tarko, L. *Rev.Chim.(Bucuresti)* **1999**, *50,* 864.
4. Burghelea, T.; Putina, G.; Tarko, L. *Sinteza bazata pe studiu QSAR pentru noi compusi din clasa medicamentelor β-blocante,* Cercetarea medicamentului intre informatie si stiintele vietii (ed. II) - Bucuresti-Romania, 1999.
5. Guta, R.; Ilie, C.; Andreescu, D. N.; Ghita, C.; Surmeian, M.; Sever, S.; Tarko, L. *Noi compusi de tip 1,4 dihidropiridinic (esteri asimetrici) cu potentiala actiune biologica,* Cercetarea medicamentului intre informatie si stiintele vietii (ed. II) – Bucuresti-Romania, 1999.
6. Croitoru, M.; Ion, I.; Dinca, A.; Rughinis, D.; Tarko, L. *Sinteza unor compusi noi, potential biologic activi, din clasa acizilor N-arilantranilici, (Poster nr. 10),* Cercetarea medicamentului intre informatie si stiintele vietii (ed. II) – Bucuresti-Romania, 1999.
7. Tarko, L. *Rev. Roum. Chim.* **2000**, *45*, 809.
8. Tarko, L.; Ivanciuc, O. *MATC* **2001**, *44*, 201.
9. Tarko, L.; Filip, P. *Rev. Roum. Chim,* **2003**, *48*, 745.
10. Tarko, L. *Rev.Chim. (Bucuresti)* **2004**, *55*, 169.
11. Beteringhe, A.; Filip, P.; Tarko L. *ARKIVOC* **2005,** *(x)*, 45.
12. Tarko, L. *Rev.Chim. (Bucuresti)* **2005**, *56*. In press.
13. Gilbert, K.; Gajewski, J.J. *Serena Software*, Box 3076, Bloomington, IN, USA.
14. Saunders, M.; Houk, K. N.; Wu, Y.-D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. *J. Am. Chem. Soc.* **1990**, *112*, 1419
15. Stewart, J.J.P. *QCMP175*, software MOPAC 7.0.
16. Barker, J. S.; Hattotuwagama, C. K.; Drew, M.G.B. *Pure Appl.Chem.* **2002**, *74(7)*, 1207.
17. Pietrzycki, W. *Polish J.Chem.* **2001**, *75,* 1569.
18. Mazur, R. H.; Goldkamp, A. H.; James, P. A.; Schlatter, J. M. *J. Med. Chem.* **1970,** *13*, 1217.
19. Mazur, R. H.; Reuter, J. A.; Swiatek, K. A.; Schlatter, J. M. *J. Med. Chem.* **1973**, *16*, 1284.
20. MacDonald, S. A.; Willson, C. G.; Chorev, M.; Vernacchia, F. S. Goodman, M. *J.Med.Chem.* **1980**, *23*, 413.
21. Iwamura, H. *J.Med.Chem.* **1981**, *24*, 572.
22. Mapelli, C.; Newton, M. G.; Ringold, C.E.; Stammer, C. H. *Int .J. Pept. Protein Res.* **1984**, *30*, 498.
23. Janusz, J. M.; Young, P. A.; Blum, R. B.; Riley, C. M. *J.Med.Chem.* **1990**, *33*, 1052.
24. Prakash, I.; Chapeau, M.-C.D. *US Patent 6077962*, 2000.
25. Prakash, I.; Guo Zhi, *US Patent 6146680*, 2000.
26. Nofre, C.; Tinti, J.-M. *EP 759030 B1*, 1998.

27. Takemoto, T.; Amino, Y., Nakamura, R., *EP 691346*, 1996.
28. Nofre, C.; Tinti, J.-M. *US Patent 5777159*, 1998.
29. Kawai, M.; Nyfeler, R.; Berman, J. M.; Goodman, M. *J.Med.Chem.* **1982**, *25*, 397.
30. Goodman, W. M.; Mattern, R. H.; Ganzel, P.; Iacovino, R.; Saviano, M.; Benedetti, E.; *J.Peptide Science* **1998**, *4*, 229.
31. Goodman, M.; Del Valle, J. R.; Amino, Y.; Benedetti, E.; *Pure Appl.Chem.* **2002**, *74*, 1109.
32. PRECLAV software is available from Center of Organic Chemistry (CCO)-Bucharest, Romanian Academy; Director - dr.ing. Petru Filip, pfilip@cco.ro.
33. Katritzky, A. R.; Petrukhin, R.; Perumal, S.; Karelson, M.; Prakash, I.; Desai, N. *Croat. Chem. Acta* **2002**, *75*, 475.