Biochemical and Biophysical Research Communications 432 (2013) 214–225

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

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

QSAR models for inhibitors of physiological impact of *Escherichia coli* that leads to diarrhea

Andrey A. Toropov [a,*], Alla P. Toropova [a], Emilio Benfenati [a], Giuseppina Gini [b], Danuta Leszczynska [c], Jerzy Leszczynski [d], Gilberto De Nucci [e]

[a] Istituto di Ricerche Farmacologiche Mario Negri, Laboratory of Environmental Chemistry and Toxicology, Via La Masa 19, 20156 Milano, Italy
[b] Department of Electronics and Information, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
[c] Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch Street, Jackson, MS 39217-0510, United States
[d] Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, MS 39217, United States
[e] University of Sao Paulo, 415 J M Machado Ave., 13092-108, Campinas SP, Brazil

ARTICLE INFO

Article history:
Received 28 January 2013
Available online 10 February 2013

Keywords:
QSAR
OECD principles
CORAL software
STa
Diarrhea

ABSTRACTS

Quantitative structure – activity relationships (QSARs) developed to evaluate percentage of inhibition of STa-stimulated (*Escherichia coli*) cGMP accumulation in T84 cells are calculated by the Monte Carlo method. This endpoint represents a measure of biological activity of a substance against diarrhea. Statistical quality of the developed models is quite good. The approach is tested using three random splits of data into the training and test sets. The statistical characteristics for three splits are the following: (1) $n = 20$, $r^2 = 0.7208$, $q^2 = 0.6583$, $s = 16.9$, $F = 46$ (training set); $n = 11$, $r^2 = 0.8986$, $s = 14.6$ (test set); (2) $n = 19$, $r^2 = 0.6689$, $q^2 = 0.5683$, $s = 17.6$, $F = 34$ (training set); $n = 12$, $r^2 = 0.8998$, $s = 12.1$ (test set); and (3) $n = 20$, $r^2 = 0.7141$, $q^2 = 0.6525$, $s = 14.7$, $F = 45$ (training set); $n = 11$, $r^2 = 0.8858$, $s = 19.5$ (test set). Based on the proposed here models hypothetical compounds which can be useful agents against diarrhea are suggested.

1. Introduction

Quantitative structure – activity relationships (QSARs) based on the molecular descriptors [1–3] are widely used as a tool to predict biochemical and medicinal characteristics of various compounds [4–10]. Such approaches are successfully applied not only in research but have been also broadly adopted by industry.

Diarrhea is a major health problem throughout the world. Only 12 years ago, in 2000 about 22% of all deaths of children in sub-Saharan Africa, and 23% in South Asia, were attributed to diarrhea diseases in 20 [11].

There is the compound that is a consider to be a starting substance for search of new effective inhibitors of physiological impact of *Escherichia coli* (STa) that leads to diarrhea: 5-(3-bromophenyl)-1,3-dimethyl-5,11-dihydro-1H indeno – [20,10,5,6] pyrido[2,3-d] pyrimidine-2,3,6-trione (BPIPP) [11]. *E. coli* induces diarrhea when it binds to intestinal epithelial cell membrane receptor, guanylyl cyclase type C (GC-C). This process activates the enzyme to convert guanosine triphosphate (GTP) to cyclic guanosine 30,50-monophosphate (cGMP). In turn, such reaction induces activation of a cGMP-dependent protein kinase and chloride-ion channel, cystic fibrosis transmembrane conductance regulator (CFTR). Finally, activation of CFTR triggers the flux of chloride ions into the intestinal lumen and the accumulation of water and sodium ions, thus causing diarrhea [11,12].

The experimentally defined percentage of the inhibition of accumulation in T84 cells cGMP (owing to presence of *E. coli*) is the measure of the ability of a compound to become the possible addition to therapeutic arsenal against diarrhea [11,12]. Computational studies provide useful way to propose the most promising candidates for further experimental evaluation of their biological activities, including the anti-diarrhea efficiency.

The aims of the present study are: (i) the evaluation of the CORAL software [13] as a tool of the QSAR modeling of the above-mentioned percentage of inhibition; and (ii) the theoretically aided selection of compounds which can be efficiently therapeutic agents for treatment of diarrhea (using of the CORAL models).

2. Method

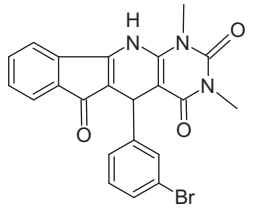
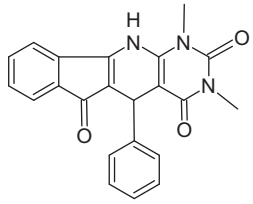
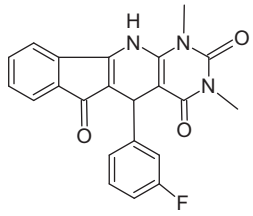
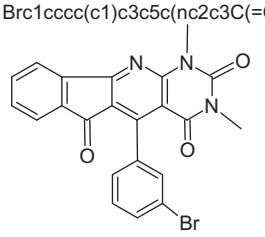
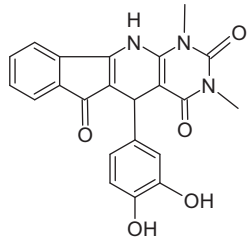
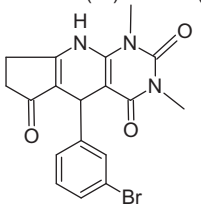
2.1. Data set

The molecular structures and percentage of inhibition of STa-stimulated cGMP accumulation in T84 cells for 31 compounds are taken from the literature [11,12]. Table 1 contains the list of

* Corresponding author.
E-mail address: andrey.toropov@marionegri.it (A.A. Toropov).

0006-291X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2013.02.011

Table 1
Inhibition of STa-stimulated cGMP accumulation in T84 cells.

ID*	SMILES and structure	% Inhibition experimental [11,12]	% Inhibition calculation with Eq. (4)
1	Brc1cccc(c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	86	51.762
2	O=C2C=1C(C4=C(NC=1c3ccccc23)N(C)C(=O)N(C)C4=O)c5ccccc5 	7	−4.210
3	Fc1cccc(c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	63	51.762
4	Brc1cccc(c1)c3c5c(nc2c3C(=O)N(C)C(=O)N2C)c4ccccc4C5=O 	2	23.709
5	Oc1ccc(cc1O)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	26	43.060
6	Brc1cccc(c1)C3C4=C(NC=2CCC(=O)C=23)N(C)C(=O)N(C)C4=O 	56	63.739

(*continued on next page*)

Table 1 (*continued*)

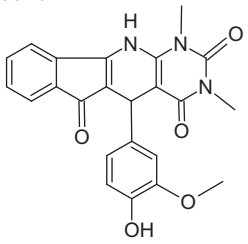
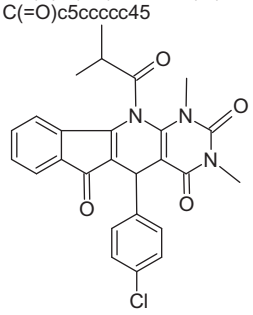
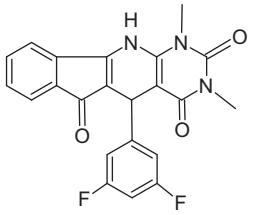
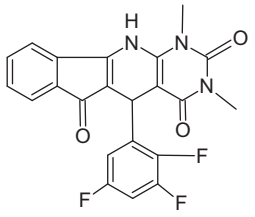
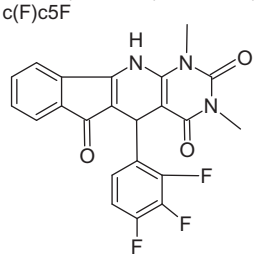
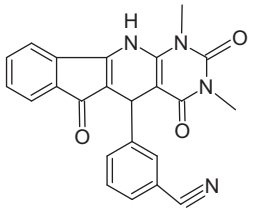
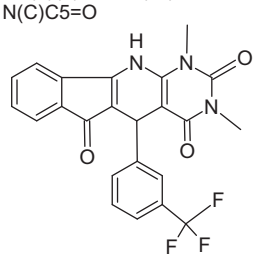
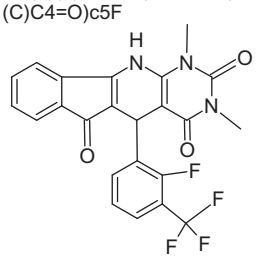
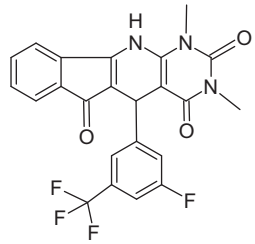
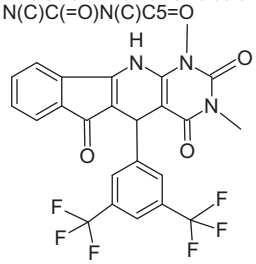
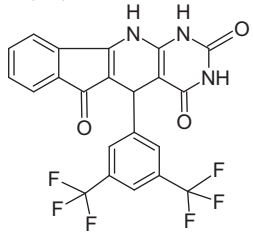
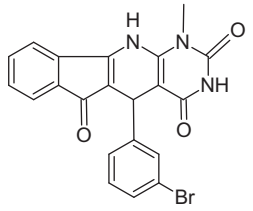
ID*	SMILES and structure	% Inhibition experimental [11,12]	% Inhibition calculation with Eq. (4)
7	Oc1ccc(cc1OC)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	29	44.848
8	CC(C)C(=O)N1C4=C(C(C2=C1N(C)C(=O)N(C)C2=O)c3ccc(Cl)cc3)C(=O)c5ccccc45 	22	23.347
17	Fc1cc(cc(F)c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	63	57.950
18	Fc5cc(C1C4=C(NC2=C1C(=O)c3ccccc23)N(C)C(=O)N(C)C4=O)c(F)c(F)c5 	42	64.138
19	Fc5ccc(C1C4=C(NC2=C1C(=O)c3ccccc23)N(C)C(=O)N(C)C4=O)c(F)c5F 	62	57.950
20	N#Cc1cccc(c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	23	55.267

Table 1 (*continued*)

ID*	SMILES and structure	% Inhibition experimental [11,12]	% Inhibition calculation with Eq. (4)
21	FC(F)(F)c1cccc(c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	73	65.927
22	FC(F)(F)c5cccc(C1C4=C(NC2=C1C(=O)c3ccccc23)N(C)C(=O)N(C)C4=O)c5F 	80	65.927
23	FC(F)(F)c1cc(cc(F)c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	93	72.115
24	FC(F)(F)c1cc(cc(c1)C(F)(F)F)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	93	86.281
25	FC(F)(F)c1cc(cc(c1)C(F)(F)F)C2C5=C(NC3=C2C(=O)c4ccccc34)NC(=O)NC5=O 	56	70.327
26	Brc1cccc(c1)C3C=5C(=O)c2ccccc2C=5NC4=C3C(=O)NC(=O)N4C 	1	−12.187

(*continued on next page*)

Table 1 (*continued*)

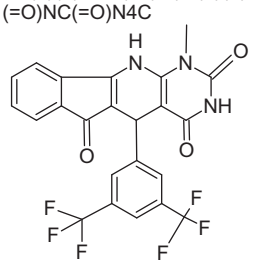
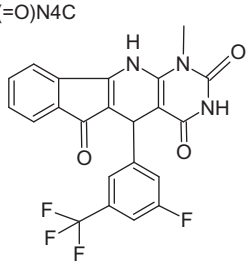
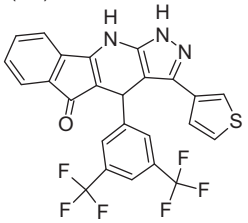
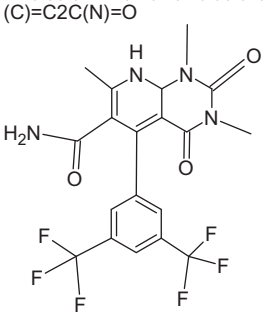
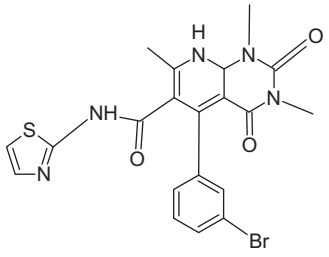
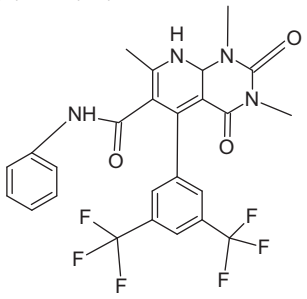
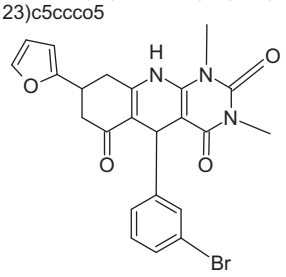
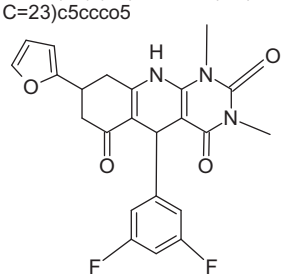
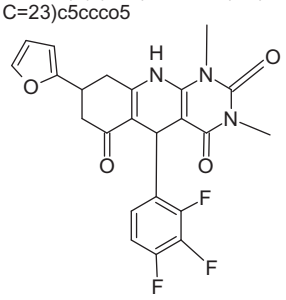
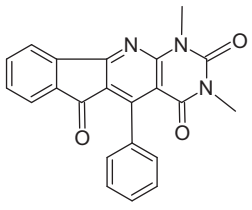
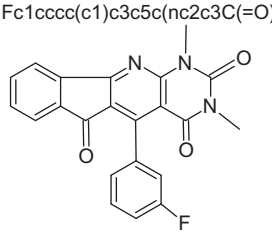
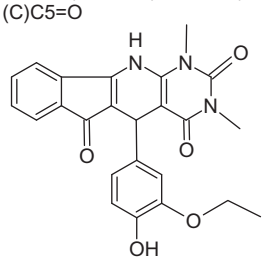
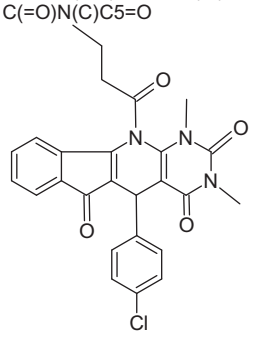
ID*	SMILES and structure	% Inhibition experimental [11,12]	% Inhibition calculation with Eq. (4)
27	FC(F)(F)c1cc(cc(c1)C(F)(F)F)C3C=5C(=O)c2ccccc2C=5NC4=C3C(=O)NC(=O)N4C 	19	22.332
28	FC(F)(F)c1cc(cc(F)c1)C3C=5C(=O)c2ccccc2C=5NC4=C3C(=O)NC(=O)N4C 	−6	8.166
29	FC(F)(F)c1cc(cc(c1)C(F)(F)F)C4c3c(nnc3c2ccsc2)NC=5c6ccccc6C(=O)C4=5 	−2	6.740
30	FC(F)(F)c1cc(cc(c1)C(F)(F)F)C2=C3C(=O)N(C)C(=O)N(C)C3NC(C)=C2C(N)=O 	8	7.502
31	O=C(Nc1nccs1)C=3C(c2cccc(Br)c2)=C4C(=O)N(C)C(=O)N(C)C4NC=3C 	11	3.392

Table 1 (*continued*)

ID*	SMILES and structure	% Inhibition experimental [11,12]	% Inhibition calculation with Eq. (4)
32	FC(F)(F)c1cc(cc(c1)C(F)(F)F)C2=C4C(=O)N(C)C(=O)N(C)C4NC(C)=C2C(=O)Nc3ccccc3 	22	21.232
33	Brc1cccc(c1)C3C=4C(=O)CC(CC=4NC=2N(C)C(=O)N(C)C(=O)C=23)c5ccco5 	2	8.058
34	Fc1cc(cc(F)c1)C3C=4C(=O)CC(CC=4NC=2N(C)C(=O)N(C)C(=O)C=23)c5ccco5 	−16	14.246
35	Fc1ccc(c(F)c1F)C3C=4C(=O)CC(CC=4NC=2N(C)C(=O)N(C)C(=O)C=23)c5ccco5 	−3	14.246
Ib	O=C5c2c(nc1c(C(=O)N(C)C(=O)N1C)c2c3ccccc3)c4ccccc45 	7	23.709

(*continued on next page*)

Table 1 (*continued*)

ID*	SMILES and structure	% Inhibition experimental [11,12]	% Inhibition calculation with Eq. (4)
IIIb	Fc1cccc(c1)c3c5c(nc2c3C(=O)N(C)C(=O)N2C)c4ccccc4C5=O 	17	23.709
VIa	Oc1ccc(cc1OCC)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	42	46.637
VII	Clc1ccc(cc1)C2C5=C(N(C3=C2C(=O)c4ccccc34)C(=O)CCC)N(C)C(=O)N(C)C5=O 	38	17.159

* ID for compounds taken from the literature [11,12].

Table 2
The identity* obtained for three random splits.

	Split 1	Split 2	Split 3
Split 1	100%	45.2%	35.5%
Split 2		100%	41.9%
Split 3			100%

* $Identity_{i,j} = \frac{E_{i,j}}{31} \times 100\%$ where $E_{i,j}$ is the number of compounds which have the same distribution (i.e. both in the training set or, vice versa, both in the test set) for i-th and j-th splits.

the considered compounds. Three random splits of the experimental data into the training set (≈67%) and test set (≈33%) are carried out. Three principles of the split are followed: (i) the range of the endpoint for the training set and test set should be as similar as possible; (ii) the splits should be random; and (iii) the splits should be different. The data displayed in Table 2 shows that three abovementioned splits are different.

2.2. Optimal descriptors

The optimal SMILES [14–17] based descriptors used to model the % inhibition are calculated as the following:

$$DCW(Threshold, N_{epoch}) = \sum CW(S_k) \quad (1)$$

where S_k is SMILES atom, i.e. one (e.g. 'C', 'N', ' = ', etc.) or two characters which cannot be examined separately ('Cl' and 'Br'); $CW(S_k)$ is correlation weight for S_k that is necessary to calculate with Eq. (1). The numerical values for $CW(S_k)$ are calculated with Monte Carlo method optimization which gives maximum of correlation coefficient between $DCW(Threshold, N_{epoch})$ and % inhibition for the training set. Having the data one can calculate $DCW(Threshold, N_{epoch})$ for all compounds of the training set and define the model as follows:

$$\%Inhibition = C_0 + C_1 \times DCW(Threshold, N_{epoch}) \quad (2)$$

The model calculated with Eq. (2) should be tested using the compounds of an external test set. Table 3 contains an example of calculation $DCW(12,39)$ for split 1. It should be noted that the descriptors are not 1D since they are based on data not only related to the bruto formula, but they are calculated with data related to presence of rings and double and triple covalent bonds.

3. Results and discussion

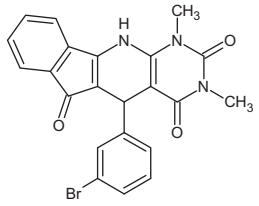
Three runs of the Monte Carlo optimization with threshold from 1 to 15 were carried out for each split. In fact, for the CORAL models the correlation coefficient between descriptor and endpoint for external test set is a mathematical function of the *Threshold* and N_{epoch}:

$$R^2_{test} = F(Threshold, N_{epoch}) \quad (3)$$

It is clear that only model with high R^2_{test} can be considered as robust according to OECD principles [18].

The performed here computational experiments have shown that preferable values of *Threshold* and N_{epoch} for various splits

Table 3
An example of DCW(Threshold, Nepoch) calculation:

Structure = 

SMILES = Brc1cccc(c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O

DCW(12,39) = 10.45450

S_k	$CW(S_k)$	N_{TRN}*	N_{TST}
Br.....	0.0	4	2
c.....	−0.9165	20	11
1.....	3.0730	20	11
c.....	−0.9165	20	11
c.....	−0.9165	20	11
c.....	−0.9165	20	11
c.....	−0.9165	20	11
(.....	0.1280	20	11
c.....	−0.9165	20	11
1.....	3.0730	20	11
(.....	0.1280	20	11
C.....	0.0740	20	11
2.....	4.1655	20	11
C.....	0.0740	20	11
5.....	1.5460	16	11
=.....	−2.0595	20	11
C.....	0.0740	20	11
(.....	0.1280	20	11
N.....	0.0710	20	11
C.....	0.0740	20	11
3.....	3.0825	20	11
=.....	−2.0595	20	11
C.....	0.0740	20	11
2.....	4.1655	20	11
C.....	0.0740	20	11
(.....	0.1280	20	11
=.....	−2.0595	20	11
O.....	−0.1800	20	11
(.....	0.1280	20	11
c.....	−0.9165	20	11
4.....	3.0335	19	11
c.....	−0.9165	20	11
c.....	−0.9165	20	11
c.....	−0.9165	20	11
c.....	−0.9165	20	11
c.....	−0.9165	20	11
3.....	3.0825	20	11
4.....	3.0335	19	11
(.....	0.1280	20	11
N.....	0.0710	20	11
(.....	0.1280	20	11
C.....	0.0740	20	11
(.....	0.1280	20	11
C.....	0.0740	20	11
(.....	0.1280	20	11
=.....	−2.0595	20	11
O.....	−0.1800	20	11
(.....	0.1280	20	11
N.....	0.0710	20	11
(.....	0.1280	20	11
C.....	0.0740	20	11
(.....	0.1280	20	11
C.....	0.0740	20	11
5.....	1.5460	16	11
=.....	−2.0595	20	11
O.....	−0.1800	20	11

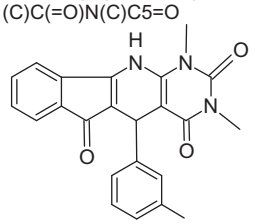
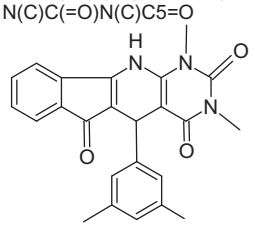
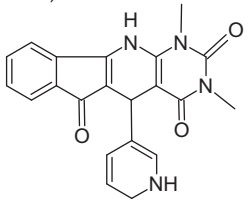
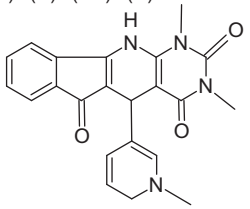
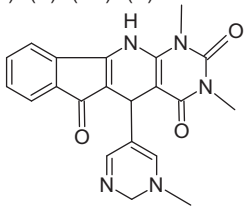
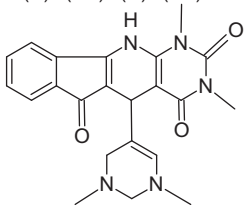
* N_{TRN} and N_{TST} are the number of a given S_k in the training and test sets, respectively.

Table 4
The numerical data on the correlation weights $CW(S_k)$ calculated by the Monte Carlo method for split 1, split 2, and split 3. Molecular features (extracted from SMILES) with stable positive values of correlation weights are indicated in bold.

No.	S_k	$CW(S_k)$ Run 1	$CW(S_k)$ Run 2	$CW(S_k)$ Run 3	N_{TRN}*	N_{TST}**
Split 1						
Promoters of increase						
1	(.....	0.09600	0.10850	0.11450	20	11
2	1.....	5.20300	4.20000	1.85100	20	11
3	2.....	2.67900	3.96650	3.34250	20	11
4	3.....	3.28850	3.72500	2.63750	20	11
5	**C.....**	**0.25300**	**0.25100**	**0.21150**	**20**	**11**
6	**N.....**	**0.04700**	**0.05750**	**0.05300**	**20**	**11**
7	4.....	3.68450	3.85100	3.72500	19	11
8	5.....	1.89700	1.97200	1.91150	16	11
Promoters of decrease						
1	=.....	−2.79500	−2.90000	−2.73850	20	11
2	O.....	−0.31350	−0.31350	−0.29150	20	11
3	c.....	−1.13350	−1.17900	−1.14050	20	11
Rare						
1	F.....	0.0	0.0	0.0	11	6
2	Br.....	0.0	0.0	0.0	4	2
3	n.....	0.0	0.0	0.0	3	2
4	Cl.....	0.0	0.0	0.0	2	0
5	**s.....**	**0.0**	**0.0**	**0.0**	**2**	**0**
6	**#.....**	**0.0**	**0.0**	**0.0**	**1**	**0**
7	6.....	0.0	0.0	0.0	1	0
8	o.....	0.0	0.0	0.0	1	2
Split 2						
Promoters of increase						
1	(.....	0.13550	0.13550	0.12500	19	12
2	1.....	3.88550	2.76550	3.22500	19	12
3	2.....	4.01450	2.93850	3.10300	19	12
4	3.....	4.33550	3.47700	4.25850	19	12
5	**C.....**	**0.23550**	**0.24150**	**0.27000**	**19**	**12**
6	**N.....**	**0.07200**	**0.12100**	**0.11650**	**19**	**12**
7	4.....	3.97900	3.91550	3.93000	18	12
8	5.....	1.86550	1.78350	1.81350	16	11
Promoters of decrease						
1	=.....	−2.89150	−2.82700	−2.89900	19	12
2	O.....	−0.35300	−0.35300	−0.37500	19	12
3	c.....	−1.18750	−1.13750	−1.15000	19	12
Rare						
1	F.....	0.0	0.0	0.0	9	8
2	Br.....	0.0	0.0	0.0	4	2
3	n.....	0.0	0.0	0.0	4	1
4	Cl.....	0.0	0.0	0.0	2	0
5	o.....	0.0	0.0	0.0	2	1
6	s.....	0.0	0.0	0.0	2	0
7	#.....	0.0	0.0	0.0	1	0
8	6.....	0.0	0.0	0.0	1	0
Split 3						
Promoters of increase						
1	(.....	0.15300	0.13850	0.13350	20	11
2	1.....	3.30850	3.41150	2.84050	20	11
3	2.....	3.36550	3.98650	2.65200	20	11
4	3.....	3.56750	3.61050	4.76150	20	11
5	**C.....**	**0.25300**	**0.28750**	**0.27700**	**20**	**11**
6	**N.....**	**0.03450**	**0.02400**	**0.02600**	**20**	**11**
7	4.....	4.46550	4.48950	4.31650	19	11
8	5.....	2.38650	2.38650	2.30400	16	11
Promoters of decrease						
1	=.....	−2.93350	−3.01250	−2.89800	20	11
2	O.....	−0.57000	−0.59900	−0.58250	20	11
3	c.....	−1.37300	−1.38550	−1.33250	20	11
Rare						
1	F.....	0.0	0.0	0.0	9	8
2	Br.....	0.0	0.0	0.0	5	1
3	n.....	0.0	0.0	0.0	5	0
4	s.....	0.0	0.0	0.0	2	0
5	#.....	0.0	0.0	0.0	1	0
6	6.....	0.0	0.0	0.0	1	0
7	Cl.....	0.0	0.0	0.0	1	1
8	o.....	0.0	0.0	0.0	1	2

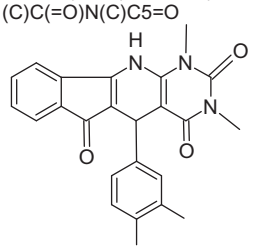
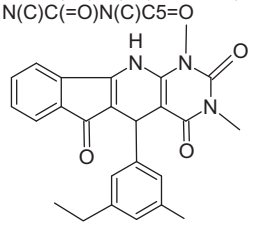
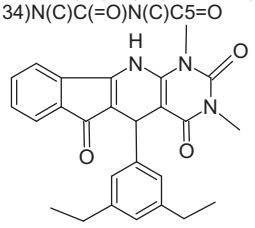
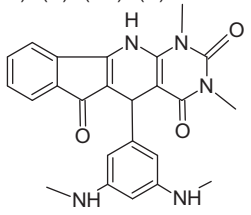
* The number of given S_k in the training set.
** The number of given S_k in the test set.

Table 5
Predicted values of % inhibition for compounds developed in this study using QSAR principles.

Probe	SMILES and structure	Eq. (4)	Eq. (5)	Eq. (6)
1	Cc1cccc(c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	53.5	51.5	53.8
2	Cc1cc(cc(C)c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	61.5	60.8	62.2
3	O=C2C=1C(C4=C(NC=1c3ccccc23)N(C)C(=O)N(C)C4=O)C=5C=CCNC=5 	−9.9	−15.8	13.0
4	CN1CC=CC(=C1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	97.5	98.2	107.8
5	CN1C=C(C=NC1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	97.4	95.3	104.2
6	CN1C=C(CN(C)C1)C3C=5C(=O)c2ccccc2C=5NC=4N(C)C(=O)N(C)C(=O)C3=4 	55.6	52.1	69.4

(*continued on next page*)

Table 5 (*continued*)

Probe	SMILES and structure	Eq. (4)	Eq. (5)	Eq. (6)
7	Cc1ccc(cc1C)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	55.3	56.0	58.0
8	Cc1cc(cc(CC)c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	63.3	65.3	66.4
9	CCc1cc(cc(CC)c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	65.1	69.7	70.5
10	CNc1cc(cc(NC)c1)C2C5=C(NC3=C2C(=O)c4ccccc34)N(C)C(=O)N(C)C5=O 	64.9	63.8	63.3

are not the same. These are *Threshold* = 12 and N_{epoch} = 39 for split 1; *Threshold* = 10 and N_{epoch} = 58 for split 2; and *Threshold* = 10 and N_{epoch} = 62 for split 3. The models obtained under these conditions are the following:

$$\%\text{inhibition} = -200.9516(\pm 7.731) + 24.1727(\pm 0.8248) * DCW(12,39) \quad (4)$$

n = 20, r^2 = 0.7208, q^2 = 0.6583, s = 16.9, F = 46 (training set) n = 11, r^2 = 0.8986, $R_m{}^2$ = 0.64, s = 14.6 (test set)

$$\%\text{inhibition} = -107.0738(\pm 4.075) + 15.9875(\pm 0.5341) * DCW(10,62) \quad (5)$$

n = 19, r^2 = 0.6689, q^2 = 0.5683, s = 17.6, F = 34 (training set) n = 12, r^2 = 0.8998, $R_m{}^2$ = 0.86, s = 12.1 (test set)

$$\%\text{inhibition} = -107.0738(\pm 4.075) + 15.9875(\pm 0.5341) * DCW(10,62) \quad (6)$$

n = 20, r^2 = 0.7141, q^2 = 0.6525, s = 14.7, F = 45 (training set) n = 11, r^2 = 0.8858, $R_m{}^2$ = 0.57, s = 19.5 (test set).

In Eqs. (4)–(6), n is the number of compounds in the set; r is correlation coefficient; s is standard error of estimation; F is Fischer F-ratio; and $R_m{}^2$ (a model has predictability if $R_m{}^2$ larger than 0.5) is metric of predictability according to Roy et al. [19–21].

One can see (Table 4) that there are three categories of molecular features: stable promoters of increase (positive correlation weights); stable promoters of decrease (negative correlation weights); and rare (correlation weights are zero). Rare molecular features for the given data are uninformative. However, two other categories of the molecular features can be used for searching compounds with the high % inhibition. One can see, that the presence of carbon atoms (sp^3) and nitrogen atoms (sp^3) should lead to increase % inhibition.

Table 5 contains some proposed by us hypothetical compounds which can be perspective efficient agents against diarrhea. Maximal inhibitor potential is revealed for probe 4 and probe 5. Per-

centage larger than 100 Eq. (6) is unfeasible, but since this is a model such values could be interpreted as "high percentage" of the inhibition.

It is necessary to follow already established rules in performed research. The OECD principles for the validation, for regulatory purposes, of (quantitative) structure–activity relationship models are the following [18]:

(1) a defined endpoint;
(2) an unambiguous algorithm;
(3) a defined domain of applicability;
(4) appropriate measures of goodness-of – fit, robustness and predictivity;
(5) a mechanistic interpretation, if possible.

In the present study endpoint represents the experimental percentage of inhibition. The algorithm is described and the CORAL software is available on the internet. The domain of applicability can be defined as the following: (i) structures which are derivatives of compound #1 (Table 1); and (ii) structures without rare molecular features (Table 4). The measures of robustness are traditional statistical characteristics: correlation coefficient (r^2) and standard error of estimation (s). Finally, the models Eqs. (4)–(6) provide possibility of classification of molecular features into three categories: promoters of increase for the endpoint, promoters of decrease of the endpoint, and rare. We believe that the data revealed by our study can be useful in order to define perspective drugs against diarrhea.

4. Conclusions

Application of CORAL software allows developing reasonable good models for the inhibition of STa-stimulated cGMP accumulation in T84 cells. The approach was tested by performing three different splits of the data into the training set and test set. The CORAL software creates models which obey the OECD principles of QSAR validation.

Acknowledgments

We thank CALEIDOS (the project number LIFE11-INV/IT 00295), and the National Science Foundation (NSF CREST HRD 0833178, and EPSCoR Award #:362492-190200-01 NSFEPS-090378) for financial support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bbrc.2013.02.011.

References

[1] I. Gutman, A.A. Toropov, A.P. Toropova, The graph of atomic orbitals and it's basic properties. 1. Wiener index // MATCH, Commun. Math. Comput. Chem. 53 (2005) 215–224.
[2] I. Gutman, B. Furtula, A.A. Toropov, A.P. Toropova, The graph of atomic orbitals and it's basic properties. 2. Zagreb index // MATCH, Commun. Math. Comput. Chem. 53 (2005) 225–230.
[3] G. Melagraki, A. Afantitis, H. Sarimveis, et al., Predictive QSAR workflow for the in silico identification and screening of novel HDAC inhibitors, Molecular Diversity 13 (3) (2009) 301–311.
[4] J. García, P.R. Duchowicz, M.F. Rozas, et al., A comparative QSAR on 1,2,5-thiadiazolidin-3-one 1,1-dioxide compounds as selective inhibitors of human serine proteinases, J. Mol. Graph. Model 31 (2011) 10–19.
[5] J.C. Garro Martinez, P.R. Duchowicz, M.R. Estrada, et al., QSAR study and molecular design of open-chain enaminones as anticonvulsant agents, Int. J. Mol. Sci. 12 (2011) 9354–9368.
[6] E. Ibezim, P.R. Duchowicz, E.V. Ortiz, et al., QSAR on aryl-piperazine derivatives with activity on malaria, Chemometr. Intell. Lab. 110 (2012) 81–88.
[7] L.M.A. Mullen, P.R. Duchowicz, E.A. Castro, QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents, Chemometr. Intell. Lab. 107 (2011) 269–275.
[8] T. Puzyn, B. Rasulev, Gajewicz, et al., Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles, Nat. Nanotechnol. 6 (2011) 175–178.
[9] A.A. Toropov, A.P. Toropova, E. Benfenati, et al., SMILES-based QSAR approaches for carcinogenicity and anticancer activity: comparison of correlation weights for identical SMILES attributes, Anti-cancer Agents Med. Chem. 11 (2011) 974–982.
[10] A.P. Toropova, A.A. Toropov, E. Benfenati, et al., CORAL: quantitative structure–activity relationship models for estimating toxicity of organic compounds in rats, J. Comput. Chem. 32 (2011) 2727–2733.
[11] E.A. Tanifum, A.Y. Kots, B.-K. Choi, et al., Novel pyridopyrimidine derivatives as inhibitors of stable toxin a (STa) induced cGMP synthesis, Bioorg. Med. Chem. Lett. 19 (2009) 3067–3071.
[12] A.Y. Kots, B.-K. Choi, M.E. Estrella-Jimenez, et al., Pyridopyrimidine derivatives as inhibitors of cyclic nucleotide synthesis: Application for treatment of diarrhea, Proc Natl Acad Sci USA 105 (2008) 8440–8445.
[13] CORAL, <http://www.insilico.eu/coral> , 2012 (accessed 15.07.12).
[14] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[15] D. Weininger, A. Weininger, J.L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, J. Chem. Inf. Comput. Sci. 29 (1989) 97–101.
[16] D. Weininger, Smiles. 3. Depict. Graphical depiction of chemical structures, J. Chem. Inf. Comput. Sci. 30 (1990) 237–243.
[17] ACD/ChemSketch Freeware, version 11.00, Advanced Chemistry Development Inc., Toronto, ON, Canada, <www.acdlabs.com>, 2007.
[18] OECD principles QSAR validation. <http://www.oecd.org/dataoecd/55/35/38130292.pdf>, 2012 (accessed August 2012).
[19] P.P. Roy, K. Roy, On some aspects of variable selection for partial least squares regression models, QSAR Comb. Sci. 27 (2008) 302–313.
[20] K. Roy, I. Mitra, On the use of the metric r_m^2 as an effective tool for validation of QSAR models in computational drug design and predictive toxicology, Mini-Rev. Med. Chem. 12 (2012) 491–504.
[21] Rm2 calculator. <http://203.200.173.43:8080/rmsquare/rmsquare-manual.html>, (accessed September 19.09.12).