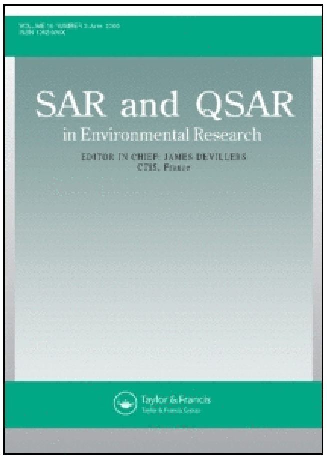


SAR and QSAR in Environmental Research

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100694

A new approach to QSAR modelling of acute toxicity

A. A. Lagunin [a]; A. V. Zakharov [a]; D. A. Filimonov [a]; V. V. Poroikov [a]
[a] Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, Russia

Online Publication Date: 01 June 2007
To cite this Article: Lagunin, A. A., Zakharov, A. V., Filimonov, D. A. and Poroikov, V. V. (2007) 'A new approach to QSAR modelling of acute toxicity ', SAR and QSAR in Environmental Research, 18:3, 285 - 298
To link to this article: DOI: 10.1080/10629360701304253
URL: http://dx.doi.org/10.1080/10629360701304253

SAR and QSAR in Environmental Research,
Vol. 18, Nos. 3–4, May–June 2007, 285–298

A new approach to QSAR modelling of acute toxicity†

A. A. LAGUNIN*, A. V. ZAKHAROV,
D. A. FILIMONOV and V. V. POROIKOV

Institute of Biomedical Chemistry, Russian Academy of Medical Sciences,
Pogodinskaya Str., 10, Moscow, 119121, Russia

(*Received 11 May 2006; in final form 20 September 2006*)

A new QSAR approach based on a Quantitative Neighbourhoods of Atoms description of molecular structures and self-consistent regression was developed. Its prediction accuracy, advantages and limitations were analysed from three sets of published experimental data on acute toxicity: 56 phenylsulfonyl carboxylates for *Vibrio fischeri*; 65 aromatic compounds for the alga *Chlorella vulgaris* and 200 phenols for the ciliated protozoan *Tetrahymena pyriformis*. According to our findings, the proposed approach provides a good correlation and prediction accuracy ($r^2 = 0.908$ and $Q^2 = 0.866$) for the set of 56 phenylsulfonyl carboxylates and the 65 aromatic compounds tested on *C. vulgaris* ($r^2 = 0.885$, $Q^2 = 0.849$). For the 200 phenols tested on *T. pyriformis*, the prediction accuracy was $r^2 = 0.685$ and $Q^2 = 0.651$. This is at least as good as the best results obtained with the other QSAR methods originally used on the same data sets.

Keywords: Acute toxicity; QSAR; *Vibrio fischeri*; *Chlorella vulgaris*; *Tetrahymena pyriformis*; QNA

1. Introduction

The QSAR studies of the environmental fate of chemicals have become a necessary tool for ecotoxicological risk assessments. Over 28 million chemicals are known to date. About 200,000 different chemicals are produced commercially and consumed every year and 2000 to 3000 new chemicals are added annually to that list. Complete toxicological data, however, are only available for less than 10,000 of them. To fill the existing data gaps, both regulatory agencies and companies have to use computational prediction models [1, 2]. Acute toxicity is one of the most important parameters in the ecotoxicological risk assessment. QSAR methods were shown to be applicable to prediction of acute toxicity [3, 4]. The pattern recognition process and multiple linear regression methods, such as MLR (multiple linear regression) and PLS

*Corresponding author. Email: alexey.lagunin@ibmc.msk.ru

†Presented at the 12th International Workshop on Quantitative Structure-Activity Relationships in Environmental Toxicology (QSAR2006), 8–12 May 2006, Lyon, France.

SAR and QSAR in Environmental Research
ISSN 1062-936X print/ISSN 1029-046X online © 2007 Taylor & Francis
DOI: 10.1080/10629360701304253

(partial least squares), are used to develop QSAR models. Key elements of QSAR modelling are training sets and molecular descriptions. The use of high quality data has became a standard in QSAR modelling. Experimental data should be taken from a well-standardized and validated assay with a clearly defined endpoint and, ideally, they should all be measured according to one and the same protocol, preferably in the same laboratory and by the same personnel [5]. In QSAR methods, different descriptions of molecules are used, including physicochemical [5–10], electro-topological, structural descriptors [11–15], etc. Each particular descriptor has its strong and weak points. There is no universal description fit for all tasks in ecotoxicological risk assessment.

We have developed a new approach for prediction of acute toxicity. It is based on a new molecular description — Quantitative Neighbourhoods of Atoms (QNA) descriptors [16] and a Self-Consistent Regression (SCR) algorithm [17]. Methods are often difficult to compare because of different numbers of compounds used in the models [5]. Therefore, for the comparison purposes, several known datasets which had been earlier used in QSAR modelling were selected.

Acute toxicity is studied on various species: unicellular organisms (e.g. algae), invertebrates (e.g. *Daphnia magna*) and vertebrates (e.g. fish). We selected three different sets (*Vibrio fischeri* [13], *Chlorella vulgaris* [9], and *Tetrahymena pyriformis* [18]) of high-quality experimental data on acute toxicity available from publications describing QSAR modelling, so that we could compare our method with the others. Comparative molecular field analysis (CoMFA) [19], multiple linear regression analysis and principal component analysis on the basis of extended topochemical atom (ETA) indices and physicochemical properties [13, 20] were used to model the *V. fischeri* acute toxicity of 56 phenylsulfonyl carboxylates. Multiple linear regression and partial least squares analyses were used to develop QSARs for *C. vulgaris* based on a small number of physicochemical descriptors [9]. A stepwise multiple regression using 108 physicochemical descriptors was applied for QSAR modelling of log (1/IGC50) for *T. pyriformis* [18]. The results of our study and the accuracy of prediction were compared with those obtained by the above-mentioned methods.

2. Materials and methods

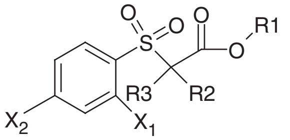
2.1 *Vibrio fischeri dataset*

In the study, we used a dataset for 56 phenylsulfonyl carboxylates with acute toxicity to *V. fischeri* (table 1). Toxicity values were presented as log EC50 (15 min-EC50, μM) [19]. Aromatic sulfones are widely used as intermediates in the production of pesticides, herbicides, and antihelmenthics, as well as floatation agents and extractants in petrochemistry and metallurgy.

2.2 *Chlorella vulgaris dataset*

The data on the *C. vulgaris* toxicity of 65 aromatic compounds [log (1/EC50), mM] were taken from Netzeva *et al*. [9]. The set included phenols, anilines, benzaldehydes, and nitrobenzenes, as well as alkyl-substituted phenols, halogenated phenols and anilines, nitro-substituted phenols, anilines and halogenated nitrobenzenes (table 2).

Table 1. Observed and predicted *V. fischeri* acute toxicity (log EC50, μM) of 56 phenylsulfonyl carboxylates.



No	*R1*	*R2*	*R3*	*X1*	*X2*	*Obs.*	*Pred.*
1	CH_3	$-(CH_2)_2-$		H	H	2.28	1.73
2	CH_3	$-(CH_2)_3-$		H	H	2.12	1.71
3	CH_3	$-(CH_2)_4-$		H	H	1.91	1.61
4	CH_3	$-(CH_2)_5-$		H	H	1.81	1.46
5	CH_3	$-(CH_2)_2-$		H	NO_2	2.12	1.52
6	$CH(CH_3)_2$	$-(CH_2)_2-$		H	NO_2	1.78	1.49
7	$CH(CH_3)_2$	$-(CH_2)_3-$		H	NO_2	1.81	1.35
8	$CH(CH_3)_2$	$-(CH_2)_5-$		H	NO_2	1.45	1.18
9	$CH(CH_3)_2$	$-(CH_2)_6-$		H	NO_2	1.05	0.89
10	CH_3	$-(CH_2)_2-$		H	Br	1.89	1.40
11	CH_3	$-(CH_2)_3-$		H	Br	1.76	1.41
12	CH_3	$-(CH_2)_4-$		H	Br	1.60	1.33
13	CH_3	$-(CH_2)_5-$		H	Br	1.31	1.20
14	CH_3	$-(CH_2)_2-$		H	Cl	1.96	1.59
15	CH_3	$-(CH_2)_3-$		H	Cl	1.92	1.54
16	$CH(CH_3)_2$	$-(CH_2)_2-$		H	Cl	1.86	1.50
17	$CH_2(CH_2)_2CH_3$	$-(CH_2)_2-$		H	Cl	1.70	1.29
18	$CH(CH_3)_2$	$-(CH_2)_4-$		H	Cl	1.51	1.16
19	$CH(CH_3)_2$	$-(CH_2)_5-$		H	Cl	1.32	0.98
20	$CH(CH_3)_2$	$-(CH_2)_6-$		H	Cl	0.90	0.80
21	$CH(CH_3)_2$	$-(CH_2)_2-$		H	CH_3	1.96	1.52
22	$CH(CH_3)_2$	$-(CH_2)_3-$		H	CH_3	1.46	1.37
23	CH_3	$-(CH_2)_2-$		H	CH_3	2.22	1.50
24	CH_2CH_3	$-(CH_2)_2-$		H	CH_3	1.92	1.56
25	CH_2CH_3	$-(CH_2)_3-$		H	CH_3	1.68	1.47
26	$CH(CH_3)_2$	$-(CH_2)_4-$		H	CH_3	1.22	1.21
27	$CH(CH_3)_2$	$-(CH_2)_5-$		H	CH_3	1.09	1.06
28	CH_3	$-(CH_2)_5-$		H	CH_3	1.40	1.11
29	CH_3	H	H	H	NO_2	1.29	0.75
30	$CH(CH_3)_2$	H	H	H	NO_2	1.28	0.95
31	CH_3	H	H	Cl	NO_2	0.44	0.42
32	$CH(CH_3)_2$	H	H	Cl	NO_2	1.13	0.74
33	CH_3	H	H	NO_2	H	1.49	0.90
34	$CH(CH_3)_2$	H	H	NO_2	H	1.34	1.09
35	CH_3	H	H	NO_2	Cl	1.33	0.66
36	$CH(CH_3)_2$	H	H	NO_2	Cl	1.45	0.91
37	CH_3	H	CH_3	H	NO_2	1.48	1.24
38	CH_3	CH_3	CH_3	H	NO_2	1.42	1.19
39	CH_3	CH_2CH_3	CH_2CH_3	H	NO_2	1.36	1.17
40	CH_3	$CH_2(CH_2)_2CH_3$	$CH_2(CH_2)_2CH_3$	H	NO_2	1.10	0.69
41	CH_3	CH_2Ph	CH_2Ph	H	NO_2	0.60	0.09
42	CH_2CH_3	$CH_2(CH_2)_2CH_3$	$CH_2(CH_2)_2CH_3$	H	NO_2	1.08	0.89
43	CH_2CH_3	CH_3	CH_2Ph	H	NO_2	0.98	0.79
44	CH_2CH_3	CH_3	$CH_2CH\_CH_2$	H	NO_2	1.12	1.07
45	CH_2CH_3	CH_3	$CH_{2}\_1\text{-Naph}$	H	NO_2	0.83	0.51
46	$CH(CH_3)_2$	$CH_2(CH_2)_2CH_3$	$CH_2(CH_2)_2CH_3$	H	NO_2	1.05	0.75
47	Cyclohexyl	H	CH_3	H	NO_2	1.19	0.94
48	CH_3	H	$CH_2CO_2CH_2CH_3$	H	NO_2	1.00	0.59
49	$CH(CH_3)_2$	H	$CH_2CO_2CH(CH_3)_2$	H	NO_2	0.92	0.57

(*Continued*)

Table 1. Continued.

No	*R₁*	*R₂*	*R₃*	*X₁*	*X₂*	*Obs.*	*Pred.*
50	$CH(CH_3)_2$	$CH_2CO_2CH_2CH_3$	$CH_2CO_2CH_2CH_3$	H	NO_2	0.66	0.45
51	CH_3		=CHPh	H	NO_2	0.82	0.31
52	CH_2CH_3		=CHPh	H	NO_2	0.75	0.39
53	$CH(CH_3)_2$		=CHPh	H	NO_2	0.64	0.20
54	$CH_2CH(CH_3)_2$		=CHPh	H	NO_2	0.66	0.50
55	$CH(CH_3)_2$		=CHPh	H	CH_3	0.89	0.38
56	$CH(CH_3)_2$		=CHPh	H	Cl	0.80	0.39
	Minimum					0.44	0.09
	Maximum					2.28	1.73

Table 2. CAS numbers, chemical names, observed and predicted *C. vulgaris* toxicity [log (1/EC50) in mM] of 65 aromatic compounds.

No	*CAS*	*Name*	*Obs.*	*Pred.*
1	108-95-2	Phenol	−1.46	−1.14
2	62-53-3	Aniline	−1.34	−0.88
3	100-66-3	Anisole	−1.09	−0.90
4	367-12-4	2-Fluorophenol	−1.08	−0.58
5	348-54-9	2-Fluoroaniline	−1.05	−0.22
6	108-39-4	3-Cresol	−1.01	−0.88
7	150-76-5	4-Methoxyphenol	−0.97	−0.89
8	95-55-6	2-Hydroxyaniline	−0.91	−0.89
9	90-05-1	2-Methoxyphenol	−0.88	−0.97
10	87-62-7	2,6-Dimethylaniline	−0.87	0.04
11	100-52-7	Benzaldehyde	−0.81	−0.91
12	95-48-7	2-Cresol	−0.81	−1.07
13	90-02-8	2-Hydroxybenzaldehyde	−0.80	−0.82
14	98-95-3	Nitrobenzene	−0.78	−0.63
15	106-44-5	4-Cresol	−0.66	−0.86
16	95-65-8	3,4-Dimethylphenol	−0.65	−0.65
17	104-87-0	4-Tolualdehyde	−0.65	−0.35
18	94-71-3	2-Ethoxyphenol	−0.62	−0.27
19	24964-64-5	3-Cyanobenzaldehyde	−0.57	−1.04
20	99-08-1	3-Nitrotoluene	−0.50	−0.08
21	106-48-9	4-Chlorophenol	−0.42	−0.34
22	97-02-9	2,4-Dinitroaniline	−0.36	0.34
23	106-41-2	4-Bromophenol	−0.35	−0.29
24	106-40-1	4-Bromoaniline	−0.33	0.19
25	108-42-9	3-Chloroaniline	−0.31	0.03
26	2495-37-6	Benzyl methacrylate	−0.21	−0.04
27	618-87-1	3,5-Dinitroaniline	0.03	0.02
28	89-98-5	2-Chlorobenzaldehyde	0.06	0.22
29	540-38-5	4-Iodophenol	0.16	−0.20
30	4748-78-1	4-Ethylbenzaldehyde	0.16	0.19
31	58-27-5	2-Methyl-1,4-naphthoquinone	0.16	0.09
32	88-69-7	2-Isopropylphenol	0.17	−0.19
33	626-43-7	3,5-Dichloroaniline	0.24	0.70
34	603-71-4	1,3,5-Trimethyl-2-nitrobenzene	0.25	0.24
35	608-31-1	2,6-Dichloroaniline	0.26	0.81
36	88-18-6	2-Tert-butylphenol	0.29	0.35
37	95-50-1	1,2-Dichlorobenzene	0.37	0.38
38	99-65-0	1,3-Dinitrobenzene	0.38	0.11
39	51-28-5	2,4-Dinitrophenol	0.40	0.32
40	100-25-4	1,4-Dinitrobenzene	0.41	0.14

(*Continued*)

Table 2. Continued.

No	*CAS*	*Name*	*Obs.*	*Pred.*
41	99-61-6	3-Nitrobenzaldehyde	0.45	0.00
42	99-30-9	2,6-Dichloro-4-nitroaniline	0.64	0.99
43	121-14-2	2,4-Dinitrotoluene	0.70	0.51
44	3531-19-9	6-Chloro-2,4-dinitroaniline	0.80	0.87
45	99-28-5	2,6-Dibromo-4-nitrophenol	0.81	1.27
46	89-61-2	2,5-Dichloronitrobenzene	0.97	1.01
47	94-62-2	Piperine	0.97	1.20
48	939-97-9	4-Tert-butylbenzaldehyde	1.00	0.76
49	634-93-5	2,4,6-Trichloroaniline	1.11	1.30
50	83-42-1	2-Chloro-6-nitrotoluene	1.17	0.78
51	5388-62-5	4-Chloro-2,6-dinitroaniline	1.19	1.16
52	528-29-0	1,2-Dinitrobenzene	1.23	0.27
53	100-00-5	1-Chloro-4-nitrobenzene	1.25	0.29
54	128-37-0	2,6-Di-tert-butyl-4-methylphenol	1.45	1.40
55	3481-20-7	2,3,5,6-Tetrachloroaniline	1.48	1.72
56	609-89-2	2,4-Dichloro-6-nitrophenol	1.50	1.12
57	83-38-5	2,6-Dichlorobenzaldehyde	1.50	0.99
58	96-76-4	2,4-Di-tert-butylphenol	1.60	1.83
59	87-86-5	Pentachlorophenol	1.69	1.86
60	89-69-0	1,2,4-Trichloro-5-nitrobenzene	1.88	1.56
61	6284-83-9	1,3,5-Trichloro-2,4-dinitrobenzene	1.89	1.96
62	1689-82-3	Phenylazophenol	2.16	1.99
63	not found	4-(Dibutylamino)benzaldehyde	2.18	2.22
64	117-18-0	2,3,5,6-Tetrachloronitrobenzene	2.34	2.22
65	608-71-9	Pentabromophenol	3.10	2.93
	Minimum		−1.46	−1.14
	Maximum		3.10	2.93

2.3 *Tetrahymena pyriformis dataset*

The data on the *T. pyriformis* toxicity of 200 phenols [log (1/IGC50), mM] were obtained from Cronin *et al.* [18]. The same compounds were selected as training set. The compounds varied in structure from phenol itself, its relatively inert alkyl and halogen derivatives, through to reactive multisubstituted phenols. Toxicity values were obtained in the population growth impairment assay on the ubiquitous freshwater ciliate *T. pyriformis* (strain GL-C). A detailed protocol can be found in Schultz [21].

2.4 *Quantitative neighbourhoods of atoms (QNA) descriptors*

The idea is that some values of each atom in a molecule can be calculated by the formula:

$$a_i \sum_k (\mathbf{f}(\mathbf{C}))_{ik} b_k, \tag{1}$$

where a_i is a property of atom i, b_k is (another) property of atom k, and $\mathbf{f}(\mathbf{C})$ is a matrix function of the molecular connectivity matrix $\mathbf{C}$. In a general case, parameters a_i and b_k may be any atomic characteristics depending on the atomic number. It is clear that each so-calculated value contains information both about the particular atom and the entire molecule. This is quite similar to a hologram, each part of which contains complete information about the whole image. In this study, we used our earlier findings [16] and, based on (1), calculated two values, P_i and Q_i, for each atom of the molecule:

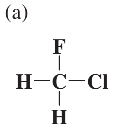
$$P_i = B_i^{-(1/2)} \sum_k \left(\exp\left(-\frac{1}{2}\mathbf{C} \right) \right)_{ik} B_k^{-(1/2)}, \tag{2}$$

$$Q_i = B_i^{-(1/2)} \sum_k \left(\exp\left(-\frac{1}{2}\mathbf{C} \right) \right)_{ik} B_k^{-(1/2)} A_k, \tag{3}$$

$$A_i = \frac{1}{2}(IP_i + EA_i), \quad B_i = IP_i - EA_i, \tag{4}$$

where IP_i is the ionisation potential (the energy required to remove the outermost electron from a neutral gaseous atom), and EA_i is the electron affinity (the energy released when an electron is added to a neutral gaseous atom of that element) of atom i (figure 1).

For the regression-based approach, however, it is essential that each object (each molecule) be presented by the same number of descriptors. This requirement poses problems when datasets contain a wide variety of structures of different size and



(b)

$$\mathbf{C} = \begin{array}{|ccccc|c} 0 & 1 & 1 & 1 & 1 & \mathbf{C} \\ 1 & 0 & 0 & 0 & 0 & \mathbf{F} \\ 1 & 0 & 0 & 0 & 0 & \mathbf{Cl} \\ 1 & 0 & 0 & 0 & 0 & \mathbf{H} \\ 1 & 0 & 0 & 0 & 0 & \mathbf{H} \end{array}$$

(c)

$$\exp(-½\mathbf{C}) = \begin{vmatrix} 1.55 & -0.59 & -0.59 & -0.59 & -0.59 \\ -0.59 & 1.14 & 0.14 & 0.14 & 0.14 \\ -0.59 & 0.14 & 1.14 & 0.14 & 0.14 \\ -0.59 & 0.14 & 0.14 & 1.14 & 0.14 \\ -0.59 & 0.14 & 0.14 & 0.14 & 1.14 \end{vmatrix}$$

(d)

	EA	IP	A = ½(IP+EA)	B=IP–EA	$B^{-½}$	P	Q
C	1.263	11.260	6.262	9.997	0.316	–0.190	–2.540
F	3.448	17.420	10.434	13.972	0.268	0.242	2.954
Cl	3.614	12.970	8.292	9.356	0.327	0.302	2.873
H	0.754	13.600	7.177	12.846	0.279	0.254	0.254
H	0.754	13.600	7.177	12.846	0.279	0.254	0.254

Figure 1. Example of QNA description for CH_2ClF: (a) structure diagram; (b) connectivity matrix; (c) exponent of connectivity matrix; (d) ionisation potentials, electron affinities, parts of equations (2) and (3), P and Q values for each atom.

different numbers of atoms. A mathematical transformation is then necessary to arrive at the same number of descriptors independent of the size of a molecule. This was done using the quantiles of QNA calculated using order statistics as follows:

$$P(F_i) = (n-1)!\sum\nolimits_k P'_k\left[\frac{F_i^{k-1}}{(k-1)!}\right]\left[\frac{(1-F_i)^{n-k}}{(n-k)!}\right],$$

$$Q(F_i) = (n-1)!\sum\nolimits_k Q'_k\left[\frac{F_i^{k-1}}{(k-1)!}\right]\left[\frac{(1-F_i)^{n-k}}{(n-k)!}\right],$$

where P'_k and Q'_k are the values of P_i and Q_i calculated according to (2)–(4) for a molecule of n atoms, and arranged in the ascending order. Choosing a certain number of values F_i, we can calculate the same fixed number of descriptors for each molecule. As a result, we present the molecular structure by the quantiles of QNA (qQNA) descriptors as a vector of the values $P(F_i)$, $Q(F_i)$, $P(F_i)Q(F_i), \dots, Q^3(F_i)$ for $i = 1, \dots, 12$ and $F_i = i/13$.

2.5 *Self-consistent regression*

Multiple Linear Regression (MLR) is based on the assumption that a response can be represented by a linear function of regressors:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{a} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is the column vector of n response values; $\mathbf{X} = (1, \boldsymbol{x}_1, \dots, \boldsymbol{x}_m)$ is the regression matrix, which consists of the column vector of units 1, and m columns of regressors $\boldsymbol{x}_k$; $\boldsymbol{a}$ is a column vector of regression coefficients should be determined; ε is a vector of unobserved residuals (errors). Residuals are usually considered as random, independent, identically distributed variables with zero mean and covariance matrix $\sigma^2\mathbf{I}_n$, where σ is the unknown standard deviation of residuals, and $\mathbf{I}_n$ is the unit matrix.

One of the best ways to solve classical MLR problems is to apply the maximum likelihood method. If the residuals $\varepsilon = \boldsymbol{y} - \mathbf{X}\boldsymbol{a}$ are normally distributed, i.e. the likelihood function $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{a})$ of the response vector $\boldsymbol{y}$ at the conditions of the regression matrix $\mathbf{X}$ and $\boldsymbol{a}$ is normal density, then the maximum likelihood method reduces to the ordinary least-squares (OLS) one:

$$\boldsymbol{a} = \arg\min \sum_i \left(y_i - \sum_k x_{ik}a_k\right)^2.$$

Although the assumption of normality is not necessary for the OLS method, it is often considered as such.

An ordinary MLR has a number of limitations. For example, the number of responses, n, should significantly exceed the number of regressors, m, and it is important to use only the non-collinear ones. A variety of techniques including stepwise and principal components regression, partial least squares, cluster significance analysis, nearest neighbour analysis and evolutionary (genetic) algorithms are used to overcome these limitations. All of such methods have certain advantages and disadvantages, the main disadvantage being their heuristic background. Another strategy employs statistical regularization of ill-posed problems [22].

It is obvious that any regressor has only a restricted influence on the response, i.e. large values of regression coefficients are prohibitive, i.e. they have small probabilities. We therefore suggest using an *a priori* probability distribution of the regression coefficients $p(\boldsymbol{a}|\boldsymbol{v})$, where $\boldsymbol{v}$ are the distribution parameters. Therefore, the estimate of $\boldsymbol{a}$ is obtained by the maximum *a posteriori* probability method:

$$\boldsymbol{a} = \arg\max p(\boldsymbol{a}|\mathbf{X}, \boldsymbol{y}, \boldsymbol{v}),$$

where $p(a|\mathbf{X}, \boldsymbol{y}, \boldsymbol{v})$ is calculated by Bayes formula:

$$p(\boldsymbol{a}|\mathbf{X}, \boldsymbol{y}, \boldsymbol{v}) = \frac{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{a})p(\boldsymbol{a}|\boldsymbol{v})}{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{v})},$$

and the likelihood function of the sample $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{v})$ is calculated by summation (integration) of all possible values of the regression coefficients $\boldsymbol{a}$:

$$p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{v}) = \sum_{a} p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{a})p(\boldsymbol{a}|\boldsymbol{v}).$$

If the residuals $\boldsymbol{\varepsilon} = \boldsymbol{y} - \mathbf{X}\boldsymbol{a}$ are normally distributed and an *a priori* conditional probability $p(\boldsymbol{a}|\boldsymbol{v})$ has also normal density:

$$p(\boldsymbol{a}|\boldsymbol{v}) \sim \exp\left[-\frac{(v_1 a_1^2 + \cdots + v_m a_m^2)}{2}\right], \tag{5}$$

then the maximum *a posteriori* probability method is the regularized least-squares method:

$$\boldsymbol{a} = \arg\min\left[\sum\nolimits_i (y_i - \sum\nolimits_k X_{ik} a_k)^2 + \sigma^2 \sum\nolimits_k v_k a_k^2\right]. \tag{6}$$

It has the following solution:

$$\boldsymbol{a} = \mathbf{T}\mathbf{X}'\boldsymbol{y}, \quad \mathrm{Var}(\boldsymbol{a}) = \sigma^2\mathbf{T}, \quad \mathbf{T} = (\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{V})^{-1}, \tag{7}$$

where $\mathbf{V}$ is the diagonal matrix of regularization parameters. In cases where $\sigma^2\mathbf{V}$ is equal to $\omega\mathbf{I}_n$ with a positive multiplier ω, equation (7) is reduced to the well-known ridge regression [23].

Based on the above, we have proposed an approach for the estimation of optimal values of parameters $\boldsymbol{v}$ in an *a priori* probability distribution $p(\boldsymbol{a}|\boldsymbol{v})$ of regression coefficients. Since the parameters $\boldsymbol{v}$ use the same data sample, $\mathbf{X}$ and $\boldsymbol{y}$, we called the method "self-consistent regression" (SCR). As recommended in [23], the maximum likelihood method can be used to find the best values of parameters $\boldsymbol{v}$:

$$\boldsymbol{a} = \arg\max\, p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{v}). \tag{8}$$

In cases where $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{a})$ and $p(\boldsymbol{a}|\boldsymbol{v})$ are the normal densities from equations (5)–(8), the following equation is derived:

$$v_k(a_k^2 + \sigma^2 t_k) = 1, \quad k = 1, \ldots, m, \tag{9}$$

where t_k is the kth diagonal element of matrix $\mathbf{T}$.

Due to their complex multidimensional nonlinear character, equation (9) can only be solved by iteration methods. Unlike the stepwise regression and other methods of combinatorial search, the SCR model includes all regressors. Nevertheless, the final model may contain several regressors truly describing the existing relationship. They can be easily identified based on their significance, which can be presented by the effective dimension of the regressor:

$$d_k = 1 - \sigma^2 v_k t_k, \quad k = 1, \dots, m. \tag{10}$$

Only those meeting a certain criterion, e.g. $d_k > 10^{-2}$, are left in the model.

The assumption of normality for $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{a})$ and $p(a|v)$ is not as restricted as seems to be the case. Normal distribution has an extreme property: it has the highest entropy for distributions with equal dispersion, and, in this sense, it is the "worst" among all possible distributions. Therefore, a solution obtained under the assumption of normality is rougher than it is theoretically possible for an exact residual distribution, but it is more robust, which is essential for the predictive power of a regression model. The regularized least-squares method (6) can be applied directly, without any statistical paradigm. However, the above-discussed statistical approach offers a useful tool for the optimisation of parameters $\boldsymbol{v}$.

If residuals' dispersion σ^2 is unknown, then the following estimate s^2 can be used:

$$s^2 = \frac{\sum_i y_i(y_i - \sum_k x_{ik} a_k)}{(n-d)}, \quad d = 1 + \sum_k d_k.$$

Based on the above-described theory, we have developed an efficient SCR algorithm. It is based on a modified Gram–Schmidt orthogonalization, which does not require the explicit inversion of a high-dimension matrix [24].

For a test molecule, the value of y can be calculated as follows:

$$y = \sum_k x_k a_k, \quad k = 1, \dots, m$$

where m is the number of regressors (qQNA) left in the equation after the SCR-part of the training procedure.

3. Results and discussion

The predicted acute toxicity values for *V. fischeri* by SCR-qQNA are given in table 1 together with the structures of compounds and observed values. We used 56 phenylsulfonyl carboxylates as training set. The model was evaluated by leave-one out cross-validation procedure (Q^2). This procedure consists in excluding the data (qQNAs and toxicity values) of each molecule (one at a time) from the training set, with the following prediction of the excluded value using the retrained model, and comparison of the predicted value and the observed one. The theorem of unbiasedness and justifiability of leave-one-out cross-validation criterion was proved by Vapnik [25]. There was a good correlation between the observed and predicted values. We use the following parameters of the regression models: n is the number of compounds in the training set, r^2 is the square of the regression coefficient and Q^2 is the cross-validated r^2, F is the Fisher's statistics, SD is a standard deviation. The statistical parameters of the correlation are $n = 56$, $r^2 = 0.908$, $Q^2 = 0.866$, $F = 58$, SD = 0.164. The final equation

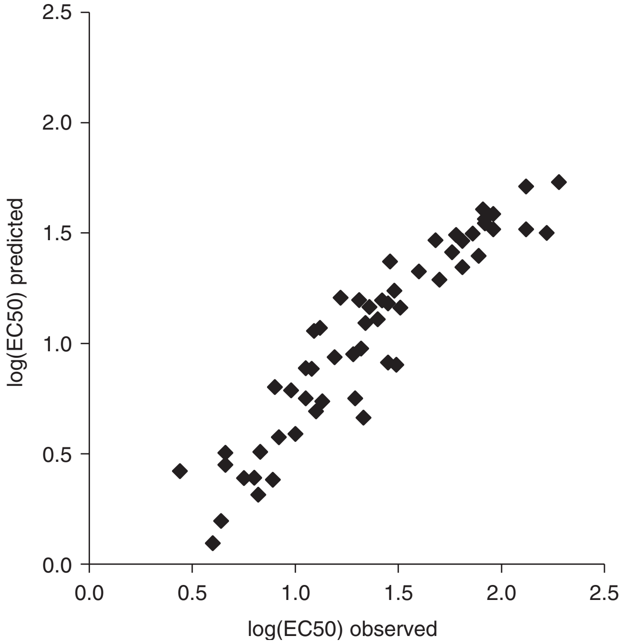


Figure 2. Toxicity to *V. fischeri*: SCR-qQNA-predicted *vs*. observed values.

contains 8 QNA quantiles as independent variables. A plot of the observed *versus* predicted by the SCR-qQNA method is presented in figure 2.

All compounds in the set are phenylsulfonyl carboxylates that are structurally similar to each other. The observed values of acute toxicity vary from 0.44 to 2.28 log EC50 (μM), while those predicted vary from 0.09 to 1.77 log EC50 (μM). For 45 compounds (80%), the deviation of the predicted values from the observed ones is less than 0.5 log EC50 (μM). A comparison of the prediction accuracy of SCR-qQNA with the other QSAR methods applied to the same data is presented in table 3.

The studied set of compounds was used for CoMFA [19], MLR analysis, principal component regression (PCR) analysis and Genetic Function Approximation (GFA) on the basis of extended topochemical atom (ETA) indices and non-ETA (physicochemical) parameters [13, 20]. Non-ETA parameters include topological indices such as Wiener, Hosoya Z, molecular connectivity, kappa shape, Balaban J and E-State parameters, as well as physicochemical parameters such as AlogP98, MolRef, and H-bond-acceptor. It is clear that the SCR-qQNA method has the best statistical parameters of correlation. The CoMFA method showed an excellent r^2 value (0.92) which was slightly better then that of the SCR-qQNA method, but Q^2 (0.79) was worse. Cross-validated Q^2 values are typically lower than normal r^2 values, yet they are considered to be more indicative of the predictive ability of the model. Whereas r^2 is a measure of goodness of fit, Q^2 is a measure of goodness of prediction. Therefore, a higher value of Q^2 is more important for prediction of toxicity than a higher value of r^2.

Factor scores were used as independent variables so as to apply the backward stepwise regression method (Factor score, PCR, MLR^c) on the basis of a combination

Table 3. Comparison of prediction accuracies of SCR-qQNA and other methods used for *V. fischeri* toxicity of 56 phenylsulfonyl carboxylates (log EC50, μM).

Method	*No*	r^2	Q^2	*F*	*SD*	*Ref.*
SCR-qQNA	56	0.908	0.866	58.0	0.164	
CoMFA	56	0.920	0.790	N/A	N/A	[17]
PCR, MLR[a]	56	0.837	0.726	57.4	0.186	[13]
PCR, MLR[b]	56	0.798	0.763	44.3	0.207	[13]
PCR, MLR[c]	56	0.798	0.763	44.3	0.207	[13]
Factor score, PCR, MLR[a]	56	0.894	0.816	57.8	0.196	[13]
Factor score, PCR, MLR[b]	56	0.871	0.828	55.4	0.190	[13]
Factor score, PCR, MLR[c]	56	0.905	0.848	77.5	0.178	[13]
GFA[a]	56	0.861	0.771	69.0	0.172	[18]
GFA[b]	56	0.820	0.805	53.5	0.196	[18]
GFA[c]	56	0.865	0.779	71.7	0.169	[18]

SCR-qQNA – Self-Consistent Regression on the basis of QNA quantiles. PCR – Principal Component Regression; MLR – Multiple Linear Regression; Factor score – factor scores were used as independent variables so that the backward stepwise regression method could be applied. GFA – Genetic Function Approximation; a – Extended topochemical atom (ETA) indices were used as parameters of the molecules [13]; b – Non-ETA parameters (topological indices including Wiener, Hosoya Z, molecular connectivity, kappa shape, Balaban J and E-State parameters, as well as physicochemical parameters such as AlogP98, MolRef and H-bond-acceptor) were used as parameters of the molecules; c – Both ETA and non-ETA descriptors were used as parameters of the molecules. N/A – Not available.

of the ETA indices and non-ETA descriptors. The obtained statistical parameters were very close to those of the SCR-qQNA method. The Fisher value alone was better than that of the SCR-qQNA method. However, the other parameters — r^2 (0.905), Q^2 (0.848), and the standard deviation (0.178) — were worse.

For *C. vulgaris*, we obtained similar results. Table 2 shows CAS Registration Numbers, chemical names, observed and predicted acute toxicity [log (1/EC50)] for 65 aromatic compounds tested on *C. vulgaris*. A reasonable correlation was obtained between the observed and predicted toxicity values. The statistical parameters of the correlation were as follows: $n = 65$, $r^2 = 0.885$, $Q^2 = 0.849$, $F = 41.8$, $\mathrm{SD} = 0.422$. The final equation contains 10 QNA quantiles as variables. A plot of the observed *versus* predicted toxicity values by the SCR-qQNA method is shown in figure 3.

The experimental values of acute toxicity vary from −1.46 to 3.10 log (1/EC50). The predicted values of acute toxicity vary from −1.14 to 2.93 log (1/EC50). It means that, unlike the predicted values for *V. fischeri*, the predicted values for *C. vulgaris* fall within the observed boundaries. A comparison of the SCR-qQNA accuracy with that of the other QSAR methods used for the same compounds is presented in table 4.

The set of compounds tested on *C. vulgaris* was used for QSAR modelling by MLR and Partial Least Squares Regression (PLS) on the basis of 102 molecular descriptors calculated by ClogP, MOPAC93, TSAR 3.3 (Oxford Molecular Limited, Oxford, England) and QSARis ver. 1.1 software (SciVision–Academic Press, San Diego, CA). MLR was made by MINITAB ver. 13.1 (Minitab Inc., State College, PA) and PLS was made with SIMCA-P ver. 9.0 (Umetrics AB, Umeå, Sweden) [9]. The statistical characteristics of our method are better than those achieved by MLR and PLS. Closest to our characteristics were those of the PLS hydrophobicity/electrophilicity model: $r^2 = 0.858$, $Q^2 = 0.843$, $\mathrm{SD} = 0.403$.

A plot of the observed toxicity values [log (1/IGC50)] *versus* those predicted by the SCR-qQNA method for 200 phenols tested on *T. pyriformis* is shown in figure 4.

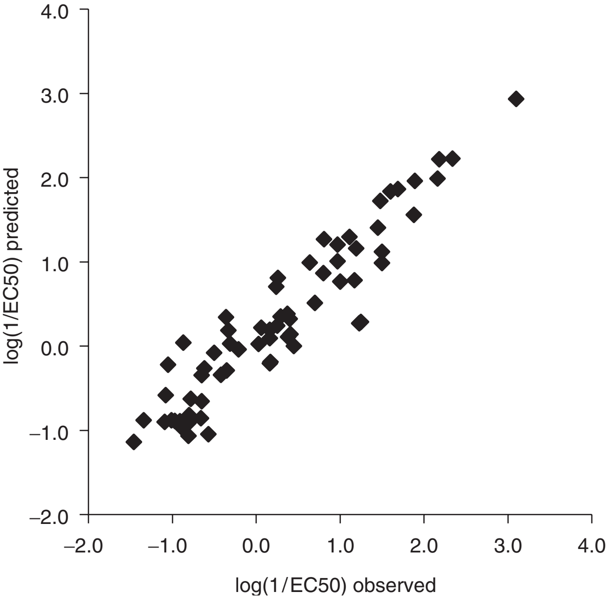


Figure 3. Toxicity to algae: SCR-qQNA-predicted *vs*. observed values.

Table 4. Comparison of prediction accuracies of SCR-qQNA and other methods for *C. vulgaris* toxicity of 65 aromatic compounds [log (1/EC50), mM].

Method	*No*	r^2	Q^2	*F*	*SD*	*Ref.*
SCR-qQNA	65	0.885	0.849	41.8	0.422	
MLR	65	0.839	0.819	161	0.429	[9]
PLS	65	0.858	0.843	N/A	0.403	[9]

SCR-qQNA – Self Consistent Regression on the basis of QNA quantiles. MLR – Multiple Linear Regression; PLS – Partial Least Squares Regression; N/A – Not available.

The statistical parameters of the correlation are as follows: $n=200$, $r^2=0.685$, $Q^2=0.651$, $F=34$, SD = 0.49. The final equation contains 12 variables.

The experimental values of acute toxicity varied from −1.5 to 2.71 [log (1/IGC50) in mM]. The predicted values of acute toxicity vary from −0.68 to 2.88 [log(1/IGC50)]. A comparison of the SCR-qQNA accuracy with that of other QSAR methods applied to the same compounds is presented in table 5.

The set of compounds tested on *T. pyriformis* was used for QSAR modelling by MLR and PLS on the basis of 108 physicochemical descriptors calculated by ACD/Labs software, Chem-X version 2000.1, MOPAC ver. 6.49, TSAR 3.3, and QSARis ver. 1.1 software. MLR was made by MINITAB ver. 13.1 and PLS was made in TSAR 3.3 [9]. It is clear that the degree of correlation between the observed and predicted values was the worst, as compared to *C. vulgaris* and *V. fischeri*. Yet, if we compare it with the other methods used to predict log (1/IGC50) for *T. pyriformis*, we can see that the statistical characteristics of our method are as good as the best ones achieved with the MLR model, which presented the following statistical parameters: $r^2=0.690$, $Q^2=0.670$, $F=75$, SD = 0.46.

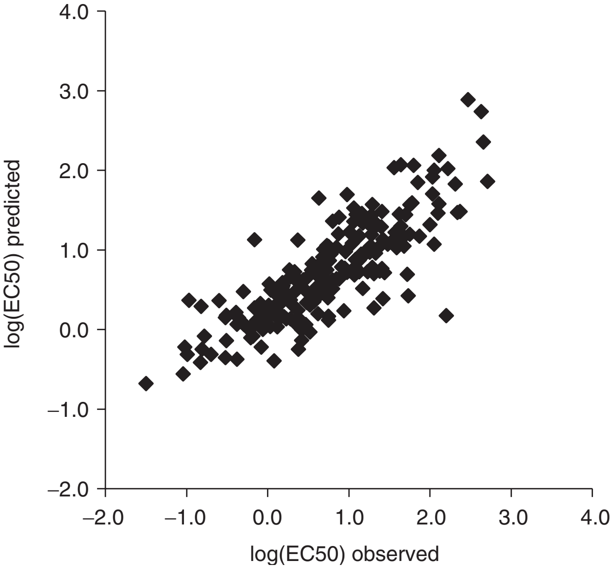


Figure 4. Toxicity to *T. pyriformis*: SCR-qQNA-predicted *vs*. observed values.

Table 5. Comparison of prediction accuracies of SCR-qQNA and other methods for *T. pyriformis* toxicity of 200 phenols [log (1/IGC50), mM].

Method	*No*	r^2	Q^2	*F*	*SD*	*Ref.*
SCR-qQNA	200	0.685	0.651	34	0.49	
MLR	200	0.690	0.670	75	0.46	[16]
PLS	200	0.600	N/A	N/A	N/A	[16]

SCR-qQNA – Self Consistent Regression on the basis of QNA quantiles; MLR – Multiple Linear Regression; PLS – Partial Least Squares Regression; N/A – Not available.

The new QSAR approach described in this paper can be used to develop a model with high correlation coefficients and low errors of prediction. Its prediction accuracy is at least as good as that of any other method used in QSAR modelling for the prediction of acute aquatic toxicity. The SCR-qQNA method is a very promising tool for prediction of the toxicity of organic compounds.

Acknowledgements

We thank the Russian Ministry of Education and Science (grants # 02.434.11.1014 and # 02.442.11.7430) for support of this work.

References

[1] M. Cronin, J. Jaworska, J. Walker, M. Comber, C. Watts, A. Worth. *Environ. Health Perspect.*, **111**, 1391 (2003).
[2] J. Walker. *J. Mol. Struct. (THEOCHEM)*, **622**, 167 (2003).
[3] M. Cronin, J. Dearden. *Quant. Struct.-Act. Relat.*, **14**, 1 (1995).

[4] M. Cronin, J. Dearden. *Quant. Struct.-Act. Relat.*, **14**, 518 (1995).

[5] M. Cronin, T. Netzeva, J. Dearden, R. Edwards, A. Worgan. *Chem. Res. Toxicol.*, **17**, 545 (2004).

[6] M. Cronin, T. Schultz. *Chem. Res. Toxicol.*, **14**, 1284 (2001).

[7] Y. Zhao, G. Ji, M. Cronin, J. Dearden. *Sci. Total Environ.*, **216**, 205 (1998).

[8] P. Von der Ohe, R. Kuhne, R. Ebert, R. Altenburger, M. Liess, G. Schuurmann. *Chem. Res. Toxicol.*, **18**, 536 (2005).

[9] T. Netzeva, J. Dearden, R. Edwards, A. Worgan, M. Cronin. *J. Chem. Inf. Comput. Sci.*, **44**, 258 (2004).

[10] T. Schultz. *Chem. Res. Toxicol.*, **12**, 1262 (1999).

[11] F. Burden, D. Winkler. *Chem. Res. Toxicol.*, **13**, 436 (2000).

[12] V. Agrawala, P. Khadikar. *Bioorg. Med. Chem.*, **10**, 3517 (2002).

[13] K. Roy, G. Ghosh. *QSAR Comb. Sci.*, **23**, 526 (2004).

[14] T. Martin, D. Young. *Chem. Res. Toxicol.*, **14**, 1378 (2001).

[15] T. Oberg. *Chem. Res. Toxicol.*, **17**, 1630 (2004).

[16] D. Filimonov, A. Lagunin, V. Poroikov. In *Proceedings of the 15th European Symposium on Structure-Activity Relationships* (*QSAR*) *and Molecular Modelling*. E. Aki, I. Yalcin (Eds), p. 98, Istanbul (2004).

[17] D. Filimonov, D. Akimov, V. Poroikov. *Pharm. Chem. J.*, **1**, 21 (2004).

[18] M. Cronin, A. Aptula, J. Duffy, T. Netzeva, P. Rowe, I. Valkova, T. Schultz. *Chemosphere*, **49**, 1201 (2002).

[19] X. Liu, Z. Yang, L. Wang. *SAR QSAR Environ. Res.*, **14**, 183 (2003).

[20] K. Roy, G. Ghosh. *Bioorg. Med. Chem.*, **13**, 1185 (2005).

[21] T. Schultz. *Toxicol. Meth.*, **7**, 289 (1997).

[22] V. Turchin, V. Kozlov, M. Malkevich. *Usp. Fiz. Nauk (Rus)*, **102**, 345 (1970).

[23] D. Hawkins, S. Basak, X. Shi. *J. Chem. Inf. Comput. Sci.*, **41**(3), 663 (2001).

[24] G.A.F. Seber. *Linear Regression Analysis*, John Wiley and Sons, New York (1977).

[25] V.N. Vapnik. *Estimation of Dependencies Based on Empirical Data*, p. 327, Springer-Verlag, New York (1982).