計算科学の手法による有機フォスフォニウム塩の毒性評価

著者	Funar-Timofei Simona, Popa Adriana, 鈴木 孝弘
雑誌名	東洋大学紀要, 自然科学篇
号	55
ページ	63-78
発行年	2011-03
URL	http://id.nii.ac.jp/1060/00000098/

Computational Methods Applied to Organic Phosphonium Salts Toxicity

Simona FUNAR-TIMOFEI,* Adriana POPA* and Takahiro SUZUKI†

Abstract

The polymer-bound phosphonium salts are used as disinfectants, antifouling coatings, and in fiber finishing, water and air disinfection. In this study the acute toxicity (expressed as the logarithm of intravenous lethal dose for mouse) was related to the structural features of a series of organic phosphonium salts. The structure of these compounds was modelled by molecular mechanics calculations and descriptors were then derived from the minimized structures. The quantitative-structure toxicity relationships were modelled by applying multiple linear regression (MLR) combined with genetic algorithm, used for variable selection. The MLR models were then compared to those obtained by principal component regression analysis. The variables included in the final MLR models, having best goodness of fitting and prediction results, were further used in principal component regression analysis. Similar statistical results with those of the best MLR model were obtained. Geometrical and electron distribution descriptors influence the compound toxicity.

Keywords: phosphonium salts, toxicity, LD_{50}, quantitative structure-toxicity relationships (QSTR), multiple linear regression (MLR), principal component regression (PCR)

1. Introduction

During the last two decades, continuous effort has been made to develop the polymers with antimicrobial function (Kenawy et al., 2002; Kanazawa et al., 1993; Kanazawa et al., 1994). Generally, polymeric antimicrobial agents have following advantages: are nonvolatile, chemically stable, and do not permeate through the skin (Dizman et al., 2004; Lenoir et al., 2006). As a result, the application of polymers with antibacterial activities will be a major step toward a healthier living.

Polyethylene glycols (PEGs) are polymers of ethylene oxide with the generalized formula $HO(CH_2CH_2O)_n$-H, "n" indicating the average number of oxyethylene groups. The polyethylene glycols (PEGs) and anionic or nonionic PEG derivatives, are used in a great

* Romanian Academy, Institute of Chemistry Timisoara, 24 M.Viteazu Av. 300223-Romania
† Natural Science Laboratory, Toyo University, 5-28-20 Hakusan, Bunkyo-ku, Tokyo 112-8606, Japan

variety of cosmetic applications because of their solubility and viscosity properties, in addition to their low toxicity. The PEGs, their ethers, and their fatty acid esters produce little or no ocular or dermal irritation and have extremely low acute and chronic toxicities (Fruijtier-Polloth, 2005).

The phosphonium salts grafted on polyethylene glycol were found to have antibacterial activity against *Staphylococcus aureus* and *Escherichia coli* (Popa et al., 2003).

Cationic polymers were obtained by the quaternization of poly(oxyethylene)s with phosphonium salts (Popa et al., 2003; Popa et al., 2002). Lethal doses of the poly(oxyethylene)s functionalized with quaternary phosphonium end groups were determined by white mice and were calculated by the Probit method. According to the toxicity scale of Hodge and Steaner they can be considered as low toxic compounds (Popa et al., 2002).

This paper presents a structure - toxicity study applied to a series of 28 phosphonium salts derivatives, by relating their structural descriptors to the logarithm of LD_{50} values (intravenous lethal dose for mouse) by multiple linear regression (MLR) and principal component regression (PCR). Descriptors chosen by the genetic algorithm were further used in principal regression analysis, to find out structural features which influence the compound toxicity.

2. Materials and Methods

Molecular descriptors

Twenty eight quaternary phosphonium salts derivatives (in Figure 1) with known toxicity, the logarithm of the lethal intravenous dose for mouse LD_{50}, expressed in mg/Kg (calculated for the cationic structures), were employed in the quantitative structure-toxicity relationships (QSTR) study. The data were retrieved from the RTECS database (RTECS Database, MDL Information Systems, Inc. 14600 Catalina Street San Leandro, California U.S.A. 94577, http://www.ntis.gov/products/types/databases/rtecs.asp) .

The molecular structure of the phosphonium salts (modeled as cations) was built by the ChemOffice package (ChemOffice 6.0, CambridgeSoft.Com, Cambridge, MA, U.S.A.) and energetically optimized using the molecular mechanics approach. Twenty-two types of descriptors were calculated by the Dragon software (Dragon Professional 5.5/2007, Talete S.R.L., Milano, Italy), like: constitutional, topological (PW5 - path/walk 5 - Randic shape index), walk and path count, connectivity indices (X1v - valence connectivity index chi-1), information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalues (BEHm3 - highest eigenvalue n. 3 of Burden matrix / weighted by atomic masses), topological charge indices, eigenvalue-based indices (VRD1 - Randic-type eigenvector-based index from distance matrix; VRp1 - Randic-type eigenvector-based index from polarizability weighted distance matrix), Randic molecular profiles (SP06 - shape profile no. 06), geometrical (QZZe - Qzz COMMA2 value / weighted by atomic Sanderson

electronegativities, ASP - asphericity), RDF descriptors (RDF045u - Radial Distribution Function - 4.5 / unweighted), 3D-MoRSE (Mor10e - 3D-MoRSE - signal 10 / weighted by atomic Sanderson electronegativities; Mor05u - 3D-MoRSE - signal 05 / unweighted; Mor30m - 3D-MoRSE - signal 30 / weighted by atomic masses), WHIM descriptors (Ks - K global shape index / weighted by atomic electrotopological states; Te - T total size index / weighted by atomic Sanderson electronegativities; Ap - A total size index / weighted by atomic polarizabilities; P2u - 2nd component shape directional WHIM index / unweighted; G3s - 3st component symmetry directional WHIM index / weighted by atomic electrotopological states; Km - K global shape index / weighted by atomic masses), Getaway descriptors (R3v+ - R maximal autocorrelation of lag 3 / weighted by atomic van der Waals volumes ; R8v+ - R maximal autocorrelation of lag 8 / weighted by atomic van der Waals volumes; R6e - R autocorrelation of lag 6 / weighted by atomic Sanderson electronegativities; R4p+ - R maximal autocorrelation of lag 4 / weighted by atomic polarizabilities; R1u - R autocorrelation of lag 1 / unweighted ; R6u - R autocorrelation of lag 6 / unweighted; R3m - R autocorrelation of lag 3 / weighted by atomic masses; R5m - R autocorrelation of lag 5 / weighted by atomic masses R1e - R autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities; H1u - H autocorrelation of lag 1 / unweighted; HATS6u - leverage-weighted autocorrelation of lag 6 / unweighted; HATS6m - leverage-weighted autocorrelation of lag 6 / weighted by atomic masses; H6v - H autocorrelation of lag 6 / weighted by atomic van der Waals volumes; HATS3v - leverage-weighted autocorrelation of lag 3 / weighted by atomic van der Waals volumes; HATSp - leverage-weighted total index / weighted by atomic polarizabilities), functional group counts (nCp - number of terminal primary C(sp3)), atom-centred fragments (C-024 - R–CH–R), charge, molecular properties, 2D binary fingerprints, and 2D frequency fingerprints.

Multiple linear regression (MLR)

Multiple linear regression (Wold and Dunn, 1983) calculations were performed by the STATISTICA (STATISTICA 7.1, Tulsa, StatSoft Inc, OK, USA) and MobyDigs (Todeschini et al., 2003) programs.

All the statistical tests were performed at a significance level of 5 % or less. Outliers were tested by estimating the standardized residuals of less than 3 standard deviation units (Todeschini and Consonni, 2000), as implemented in the MobyDigs program (Todeschini et al., 2003).

The goodness of prediction of the MLR models was checked by the Akaike Information Criterion (AIC) (Gentleman and Wilk, 1975), the multivariate K correlation index (KX-the multivariate correlation index of the matrix of X descriptors and KXY - the multivariate correlation index of the matrix of X descriptors and Y response variable) (Todeschini et al., 1998), Y-scrambling ($r^2_{Y-scrambling}$ and $q^2_{Y-scrambling}$) (Lindgren et al., 1996), and bootstrapping (Efron, 1982) and external validation (q^2_{ext}) parameters. All these calculations were performed by the

MobyDigs software. The leave-one out cross-validation (q^2_{LOO}) procedure (Efron , 1983; Stone, 1974; Wold, 1978) was, also, employed for the internal validation of models.

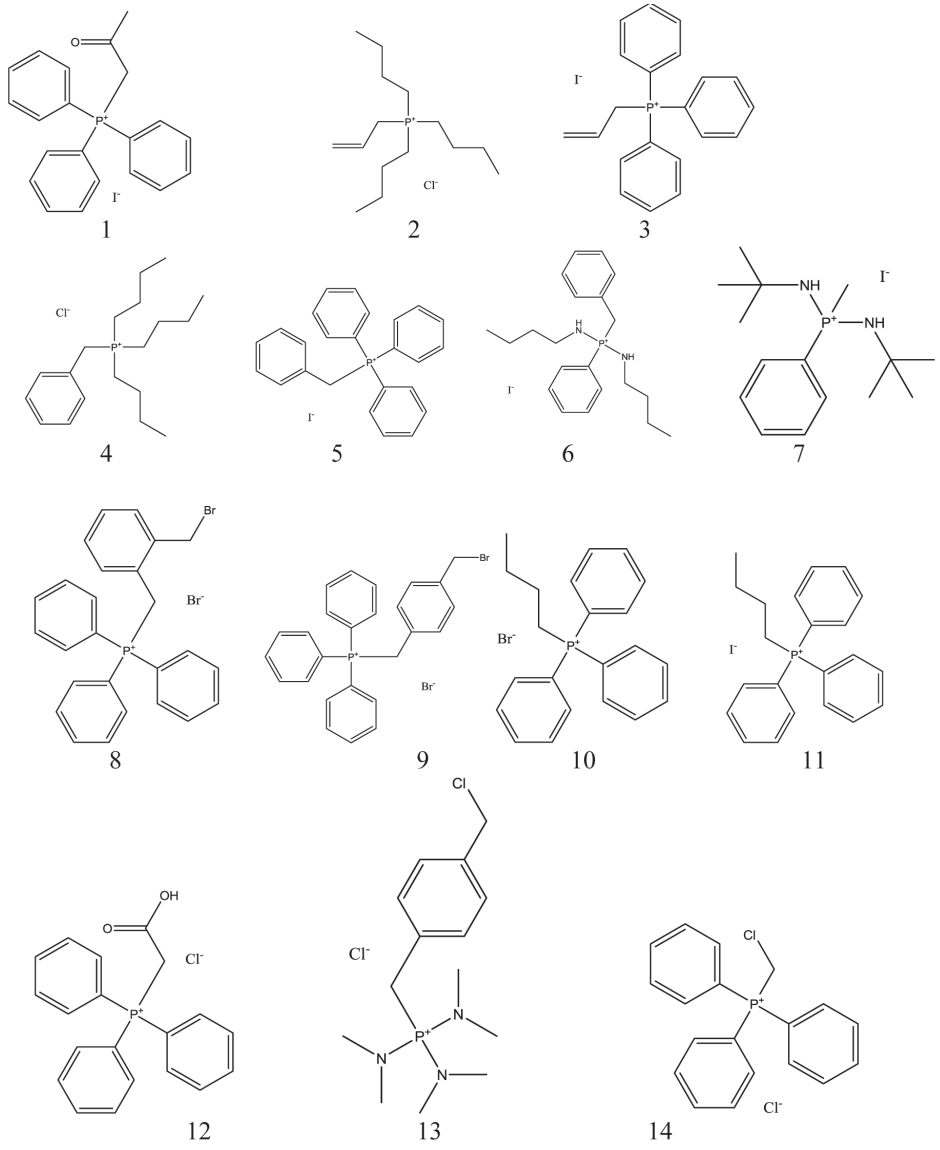


Fig 1. Phosphonium salt structure

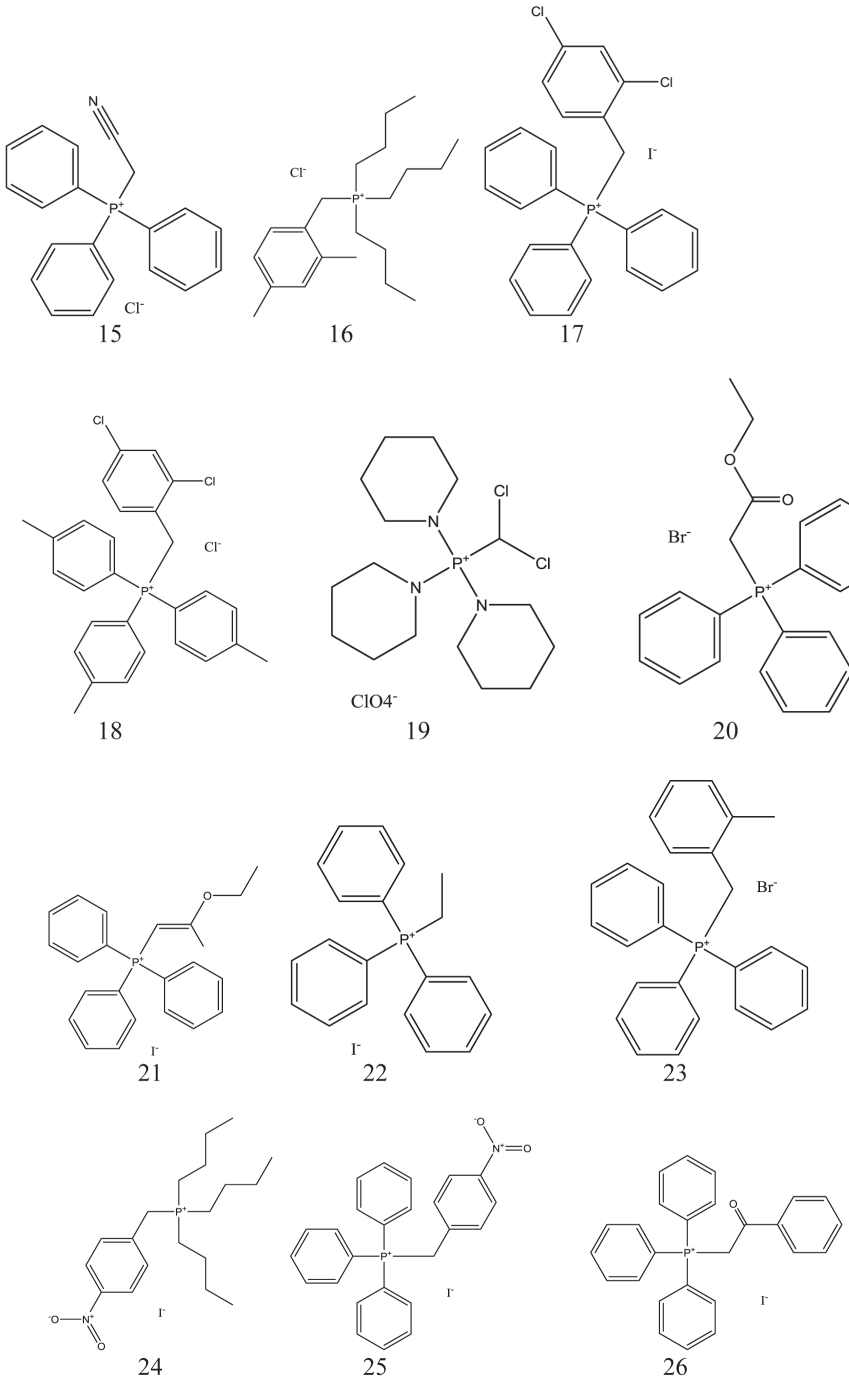
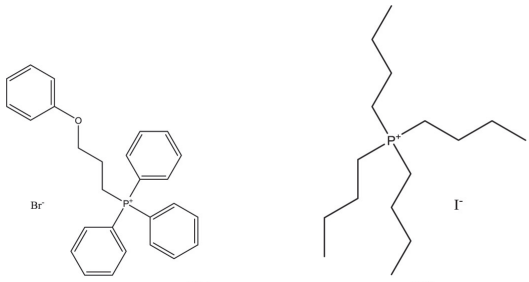


Fig 1. (continued)



27 28

Fig 1. (continued)

Principal component regression (PCR)

Principal component analysis (PCA) method (Mager and Rothe, 1990; Mager, 1994) was applied to produce new variables created by a linear combination of the original variables or descriptors. This method is based upon a spectral decomposition of the correlation matrix (H) of the regressors X_i (i = 1,2,...,c) (the physicochemical parameters). On the basis of this orthogonal transformation, principal components (PCs) of the original variables are calculated and used instead of the original variables. In PCR the multiple linear regression is applied to estimate the regression matrix of the regressands Y_i (i = 1,2,...,N) and component regressors Z_f (f = 1,2,..., s ≤ c) and finally, a re-transformation (rectification) yields the desired estimators (nonleast-squares regression). The general PCR equation is given by:

$$Y = h_o + Z \cdot H \tag{1}$$

where $h_o = \bar{y}$ is the intercept vector which is equal to the vector $\bar{y}$ of mean values of the regressands Y_i, and Z is the matrix of the principal components. The PCR calculations were performed by the MobyDigs package (Todeschini et al., 2003).

The overall performance of MLR and PCR models was evaluated in terms of the root-mean-square error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i^{fit})^2}{n}} \tag{2}$$

where y_i is the experimental target value for the ith compound ($logLD_{50}$) and y_i^{fit} represents the predicted target values calculated by the models.

3. Results and Discussion

In the MLR analysis a training set of 21 compounds and a test set of the following 7 compounds (selected randomly): 3, 4, 9, 17, 29, 30 and 32 were considered. Starting from the total set of calculated descriptors, variable selection was carried out by the genetic algorithm included in the MobyDigs program, using the RQK fitness function (Todeschini et al., 2004), with leave-one-out crossvalidation correlation coefficient as constrained function to be optimised, a crossover/mutation trade-off parameter T = 0.5 and a model population size P = 50. In addition, AIC - Akaike Information Criterion, the multivariate K correlation index (Kx and Kxy), Y-scrambling variables ($r^2_{Y-scrambling}$ and $q^2_{Y-scrambling}$), external q^2 (q^2_{ext}) values, q^2_{boot} - bootstrapping parameter and were calculated. RMSE values were calculated for the training set (SDEC values) and test set (SDEP values). The final most stable MLR models are presented in Table 2.

Starting from the descriptor matrix containing all variables, following descriptors were found to be significant and were included in the final MLR models: PW5 - path/walk 5 - Randic shape index; SP06 - shape profile no. 06; QZZe - Qzz COMMA2 value / weighted by atomic Sanderson electronegativities; ASP – asphericity; RDF045u - Radial Distribution Function - 4.5 / unweighted; Mor10e - 3D-MoRSE - signal 10 / weighted by atomic Sanderson electronegativities; Mor05u - 3D-MoRSE - signal 05 / unweighted; Mor30m - 3D-MoRSE - signal 30 / weighted by atomic masses; Ks - K global shape index / weighted by atomic electrotopological states; Te - T total size index / weighted by atomic Sanderson electronegativities; Ap - A total size index / weighted by atomic polarizabilities; R8v+ - R maximal autocorrelation of lag 8 / weighted by atomic van der Waals volumes; R6u - R autocorrelation of lag 6 / unweighted; R3m - R autocorrelation of lag 3 / weighted by atomic masses; R5m - R autocorrelation of lag 5 / weighted by atomic masses; R1e - R autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities; nCp - number of terminal primary C(sp^3); C-024 - R–CH–R.

Acceptable correlations with the phosphonium salt toxicity and models with predictive power were obtained (Table 2). The best externally predictive single model, based on four descriptors, would be selected from the population of 10 models. Considering these models, the range of q^2_{LOO} is 0.767 – 0.813, while the range of q^2_{ext} is 0.558 – 0.913. The selection criterion used in this study is that the model should have higher r^2, higher cross-validated q^2_{LOO}, higher external predictive ability, least difference between internal and external predictive ability, the fewer chemicals outside the chemical domain and the fewer chemicals with large relative errors.

Table 1. The logarithm of the lethal intravenous dose for mouse LD_{50} and descriptors included in the final MLR and PCR models*

No.	$logLD_{50}$	PW5	SP06	QZZe	ASP	RDF045u	Mor10e	Mor05u	Mor30m	Ks	Te
1	-3.76	0.13	7.03	465.74	0.04	18.41	-0.47	-5.18	0.38	0.22	14.68
2	-4.13	0.11	7.08	619.62	0.09	24.94	-0.66	-8.16	-0.18	0.29	15.62
3	-3.73	0.13	6.99	435.86	0.03	17.71	-1.00	-4.40	0.45	0.19	14.47
4	-4.47	0.11	7.64	720.25	0.24	32.34	-1.07	-9.63	-0.10	0.51	16.07
5	-3.80	0.13	7.51	552.17	0.07	26.84	-1.22	-5.66	0.35	0.26	15.41
6	-3.28	0.10	8.26	903.13	0.17	22.99	-0.74	-9.16	0.23	0.43	19.27
7	-5.68	0.08	5.66	343.46	0.09	25.05	-1.26	-3.64	0.06	0.35	10.14
8	-4.14	0.12	7.91	604.93	0.18	26.52	-1.21	-6.03	0.56	0.24	16.46
9	-4.39	0.12	8.20	698.10	0.34	26.92	-1.14	-5.80	0.63	0.39	17.44
10	-3.76	0.13	7.30	534.95	0.03	20.73	-1.07	-6.18	0.43	0.16	15.67
11	-3.25	0.13	7.30	534.95	0.03	20.73	-1.07	-6.18	0.43	0.16	15.67
12	-3.25	0.13	7.00	457.62	0.05	19.43	-0.60	-4.29	0.43	0.19	13.96
13	-6.08	0.07	7.18	477.43	0.49	31.61	-1.11	-4.65	-0.18	0.66	13.09
14	-3.99	0.12	6.89	436.37	0.05	16.25	-0.45	-4.10	0.49	0.29	13.85
15	-3.73	0.13	6.97	441.84	0.07	17.98	-0.90	-4.49	0.47	0.24	13.94
16	-4.25	0.11	7.94	895.71	0.26	37.96	-1.87	-10.19	-0.15	0.53	18.05
17	-4.37	0.12	7.93	588.00	0.14	25.69	-0.97	-5.45	0.38	0.33	16.14
18	-3.92	0.12	8.42	915.70	0.10	31.31	-2.81	-7.02	0.32	0.27	20.62
19	-4.31	0.12	6.46	519.41	0.01	35.12	-1.59	-5.51	0.19	0.11	12.33
20	-4.15	0.12	7.30	556.40	0.07	35.09	-1.62	-6.06	0.51	0.22	14.91
21	-3.79	0.12	7.71	620.75	0.04	20.82	-1.34	-5.59	0.47	0.19	17.17
22	-4.72	0.12	6.86	428.79	0.07	17.66	-0.61	-3.61	0.42	0.28	13.83
23	-4.06	0.12	7.61	588.48	0.05	26.76	-1.52	-5.50	0.50	0.24	16.18
24	-4.53	0.10	8.32	846.16	0.37	34.43	-1.41	-9.52	0.02	0.67	17.55
25	-4.35	0.12	8.29	683.83	0.21	27.54	-1.07	-5.36	0.62	0.51	17.25
26	-3.83	0.12	7.41	574.88	0.10	29.94	-1.34	-4.07	0.62	0.28	14.23
27	-3.85	0.11	9.35	1023.65	0.27	28.03	-1.60	-7.34	0.41	0.52	22.92
28	-3.91	0.12	7.29	681.39	0.07	27.17	-0.82	-8.73	-0.12	0.25	16.42

Table 1. (continued)

No.	Ap	R8v+	R6u	R3m	R5m	R1e	nCp	C-024
1	53.03	0.01	1.71	0.40	0.54	2.10	1.00	15.00
2	58.86	0.01	1.55	0.19	0.22	2.02	3.00	0.00
3	51.67	0.01	1.73	0.41	0.49	2.08	0.00	15.00
4	56.12	0.02	1.61	0.31	0.27	2.05	3.00	5.00
5	57.91	0.01	1.63	0.47	0.51	2.11	0.00	20.00
6	87.99	0.02	1.24	0.37	0.30	2.11	2.00	10.00
7	26.10	0.01	2.29	0.49	0.34	2.00	6.00	5.00
8	73.18	0.02	1.53	0.72	0.62	2.10	1.00	19.00
9	72.55	0.01	1.49	0.68	0.53	2.10	1.00	19.00
10	61.28	0.01	1.65	0.38	0.45	2.11	1.00	15.00
11	61.28	0.01	1.65	0.38	0.45	2.11	1.00	15.00
12	45.87	0.01	1.84	0.49	0.65	2.12	0.00	15.00
13	36.88	0.01	2.64	0.56	0.32	1.98	1.00	4.00
14	44.49	0.01	1.59	0.58	0.76	2.01	0.00	15.00
15	45.20	0.01	1.72	0.51	0.59	2.05	0.00	15.00
16	67.89	0.01	1.56	0.30	0.25	2.06	5.00	3.00
17	67.28	0.01	1.58	0.71	0.58	2.09	0.00	18.00
18	106.48	0.01	1.06	0.66	0.41	2.11	3.00	15.00
19	41.75	0.01	2.29	0.72	0.93	2.03	0.00	0.00
20	53.09	0.01	1.70	0.51	0.58	2.10	1.00	15.00
21	72.06	0.01	1.50	0.41	0.46	2.14	2.00	15.00
22	46.27	0.01	1.95	0.44	0.47	2.07	1.00	15.00
23	64.92	0.01	1.54	0.46	0.48	2.13	1.00	19.00
24	61.12	0.01	1.53	0.43	0.30	2.11	3.00	4.00
25	65.60	0.01	1.52	0.59	0.51	2.18	0.00	19.00
26	46.10	0.01	1.73	0.58	0.62	2.09	0.00	20.00
27	117.03	0.01	1.34	0.49	0.38	2.18	0.00	20.00
28	68.36	0.02	1.44	0.18	0.20	2.06	4.00	0.00

* PW5 - path/walk 5 - Randic shape index; SP06 - shape profile no. 06; QZZe - Qzz COMMA2 value / weighted by atomic Sanderson electronegativities; ASP – asphericity; RDF045u - Radial Distribution Function - 4.5 / unweighted; Mor10e - 3D-MoRSE - signal 10 / weighted by atomic Sanderson electronegativities; Mor05u - 3D-MoRSE - signal 05 / unweighted; Mor30m - 3D-MoRSE - signal 30 / weighted by atomic masses; Ks - K global shape index / weighted by atomic electrotopological states; Te - T total size index / weighted by atomic Sanderson electronegativities; Ap - A total size index / weighted by atomic polarizabilities; R8v+ - R maximal autocorrelation of lag 8 / weighted by atomic van der Waals volumes; R6u - R autocorrelation of lag 6 / unweighted; R3m - R autocorrelation of lag 3 / weighted by atomic masses; R5m - R autocorrelation of lag 5 / weighted by atomic masses; R1e - R autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities; nCp - number of terminal primary C(sp3); C-024 - R--CH--R

Table 2. MLR results for the series of 28 phosphonium salts*

No	Descriptors	r^2	q^2_{LOO}	q^2_{boot}	q^2_{ext}	$r^2_{Y-scrambling}$	$q^2_{Y-scrambling}$	AIC	Kx	Kxy	SDEP	SDEC	F	s	Obs.
1	QZZe Mor10e Ks R8v+	0.872	0.813	0.779	0.558	0.383	-0.019	0.127	0.36	0.43	0.295	0.244	27.33	0.279	Compound no. 8 – outlier**
2	QZZe RDF045u Ks R8v+	0.865	0.795	0.742	0.636	0.316	-0.35	0.135	0.37	0.47	0.309	0.251	25.55	0.288	Compound no. 8 – outlier**
3	R6u R3m R5m R1e	0.869	0.784	0.752	0.843	0.398	-0.248	0.13	0.45	0.51	0.317	0.247	26.5	0.283	
4	PW5 QZZe Mor10e Ks	0.870	0.779	0.776	0.845	0.57	0.062	0.129	0.44	0.52	0.321	0.246	26.77	0.282	Compound no. 6 – outlier**
5	SP06 RDF045u Mor05u Ks	0.850	0.772	0.737	0.599	0.658	0.393	0.149	0.46	0.50	0.326	0.264	22.66	0.303	Compound no. 26 – leverage point**
6	Mor10e Te Ap Ks	0.857	0.771	0.671	0.913	0.361	-0.506	0.142	0.46	0.52	0.327	0.258	23.9	0.296	Intercorrelated descriptors (Te and Ap)
7	ASP Mor05u Mor30m nCp	0.837	0.770	0.705	0.850	0.422	-0.369	0.162	0.40	0.44	0.327	0.275	20.59	0.315	Compound no. 7 – leverage point Compound no. 12 – outlier**
8	ASP Mor05u nCp C-024	0.837	0.769	0.692	0.836	0.365	-0.495	0.162	0.35	0.39	0.328	0.275	20.57	0.316	Compound no. 7, 13 – leverage points Compound no. 12 – outlier**
9	RDF045u Mor05u Mor30m Ks	0.853	0.768	0.778	0.725	0.315	-0.166	0.146	0.48	0.49	0.329	0.262	23.16	0.3	Compound no. 12 – outlier**
10	SP06 Mor05u Mor10e Ks	0.854	0.767	0.719	0.562	0.419	-0.139	0.145	0.41	0.43	0.329	0.261	23.46	0.298	

* r^2 – correlation coefficient, SDEP – standard deviation error in prediction ($RMSE_{test}$), SDEC – standard deviation error in calculation ($RMSE_{training}$), F- Fischer test, s – standard error of estimate, AIC - Akaike Information Criterion, the multivariate K correlation index (Kx and Kxy), Y-scrambling variables ($r^2_{Y-scrambling}$ and $q^2_{Y-scrambling}$), q^2_{ext} - external q^2, q^2_{boot} - bootstrapping parameter, q^2_{LOO} - leave-one out cross-validation parameter; ** The leverage points and outliers were not omitted from the models presented

Models 3 and 10 were found as most stable, whose regression equations are:

$$\log LD_{50} = -10.2(\pm 3.89) - 1.09(\pm 0.25)R6u - 3.01(\pm 0.61)R3m + 2.92(\pm 0.49)R5m + 1.73(\pm 2.2)R1e \quad (3)$$

$n_{training} = 21,\ n_{test} = 7$

$r^2_{training} = 0.869,\ q^2_{LOO} = 0.784,\ q^2_{ext} = 0.843,\ r^2_{Y-scrambling} = 0.398,\ q^2_{Y-scrambling} = -0.248$

$K_{XY} = 0.514,\ K_X = 0.445,\ RMSE_{training} = 0.247,\ RMSE_{test} = 0.317,\ F = 25.6,\ s = 0.283$

$$\log LD_{50} = -7.44(\pm 0.69) + 0.55(\pm 0.11)SP06 - 0.17(\pm 0.05)Mor05u + 0.48(\pm 0.13)Mor10e - 4.13(\pm 0.49)Ks \quad (4)$$

$n_{training} = 21,\ n_{test} = 7$

$r^2_{training} = 0.854,\ q^2_{LOO} = 0.767,\ q^2_{ext} = 0.562,\ r^2_{Y-scrambling} = 0.419,\ q^2_{Y-scrambling} = -0.139$

$K_{XY} = 0.433,\ K_X = 0.410,\ RMSE_{training} = 0.261,\ RMSE_{test} = 0.329,\ F(4,16) = 23.46,\ s = 0.298$

From all the statistical parameters, it can be seen that the proposed models are stable, robust and predictive. Figures 1 shows the regression plot of the single model 3, respectively 10. Descriptors R6u, R3m, R5m, R1e, SP06, Mor05u, Mor10e and Ks were the most important descriptors.

Following descriptors: R6u, R3m, Mor05u and Ks yielded high toxicity values. A low toxicity of phosphonium compounds was derived by the R5m, R1e, SP06 and Mor10e descriptors. The Williams plots for models 3 and 10 (Table 2) are presented in Figure 2.

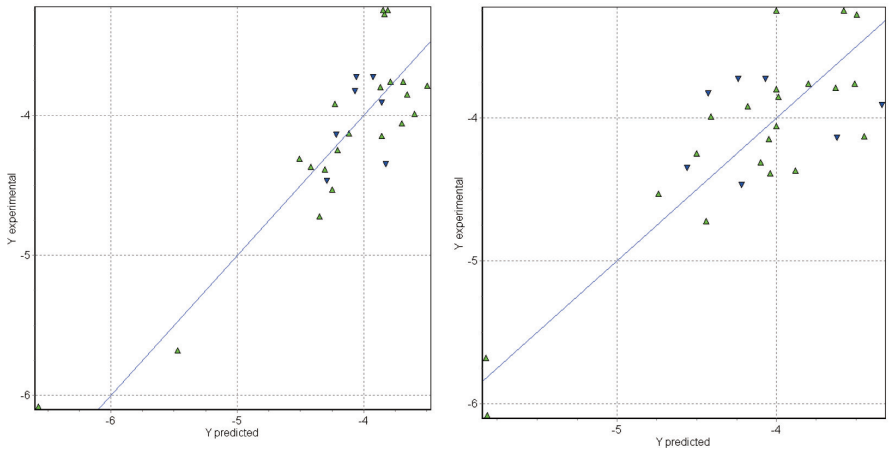


Figure 1. Experimental *versus* predicted $logLD_{50}$ values of the final MLR model 3 (left) and model 10 (right). Training set is marked by green triangles, test set marked by blue triangles.

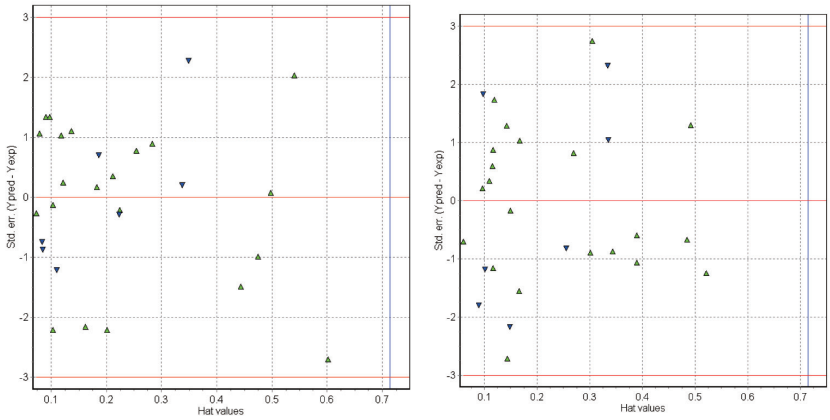


Figure 2. Williams plot: jackknifed residuals versus leverages of the MLR model 3 (left) and model 10 (right). Training set is marked by by green triangles, test set marked by blue triangles.

No outliers or influential points were found in both models 3 and 10, considered to have best statistical results.

Ks (K global shape index / weighted by atomic electrotopological states) encodes information on atomic distribution and shape (weighted by electrotopological states) along the main direction of the molecule. It was concluded that this molecule shape is very important for the phosphonium salts toxicity. WHIM descriptors are built in such a way to capture relevant molecular 3D information regarding molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames.

The SP06 (shape profile no. 06) descriptor is another parameter which expresses the molecular shape.

3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction) descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves (Schuur et al., 1996) Mor05u (3D-MoRSE - signal 05 / unweighted) and Mor10e (3D-MoRSE - signal 10 / weighted by atomic Sanderson electronegativities) descriptors are geometrical descriptors calculated for the unweighted case, respectively electronegativity.

GETAWAYs descriptors which are geometrical descriptors encoding information on the effective position of substituents and fragments in the molecular space. R6u (R autocorrelation of lag 6 / unweighted), R3m (R autocorrelation of lag 3 / weighted by atomic masses), R5m (R autocorrelation of lag 5 / weighted by atomic masses), and R1e (R autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities) are evaluated by considering separately all the contributions of each different path length (lag) in the molecular graph, as collected in the topological distance matrix. Therefore steric factors can be considered to influence the toxicity.

The PCR calculations were performed on variables chosen by the final MLR models

presented in table 2, by using the Q^2 leave-one-out as fitness parameter. The most correlated PCs to the Y response variable were selected by GA.

Five principal components were found to be significant (r^2 = 0.861, q^2_{LOO} = 0.763, $RMSE_{training}$ = 0.254, $RMSE_{test}$ = 0.332). These results are slightly poorer in comparison to the MLR model 3 ones. The PCR results are presented in table 3 and the regression coefficients in table 4. The RMSE values were of 0.215 for the training set, 0.194 for the test set, close to the RMSE values of the MLR model.

Table 3. PCR results of the model with five principal components (PCs)

Components	PC1	PC2	PC3	PC4	PC5
Eigenvalues	7.21	5.02	2.92	1.42	0.99
Exp.Var.%	38.0	26.4	15.4	7.5	5.2
X1v	0.279	-0.126	0.139	0.405	0.086
BEHm3	0.246	0.288	-0.074	-0.221	-0.066
VRD1	0.348	0.118	0.095	0.047	0.036
VRp1	0.348	0.118	0.095	0.047	0.036
Mor05e	-0.092	0.394	-0.093	-0.224	0.164
P2u	-0.081	0.112	-0.475	0.267	-0.133
G3s	-0.292	0.002	0.216	-0.017	-0.316
Km	-0.024	-0.148	0.504	0.071	-0.139
H1u	0.359	0.006	-0.083	-0.145	0.042
HATS6u	-0.288	0.258	0.014	-0.156	-0.004
HATS6m	0.076	0.336	-0.258	0.046	-0.278
H6v	0.168	0.097	0.419	-0.169	0.081
HATS3v	0.055	0.420	0.080	0.097	0.039
HATSp	0.268	0.284	0.017	0.134	-0.086
R1u	0.302	-0.150	-0.110	-0.338	0.137
R3v+	-0.046	0.302	0.341	-0.200	-0.089
R8v+	0.166	-0.103	-0.007	-0.145	-0.833
R6e	-0.274	0.206	0.099	-0.143	0.076
R4p+	-0.043	0.267	0.167	0.598	-0.030

Descriptors: X1v, BEHm3, VRD1, VRp1, G3s, H1u, HATS6u, HATSp, R1u and R6e have an important contribution to the first PC, Mor05e, HATS6m, R3v+ - to the second PC, P2u, Km, H6v, R3v+ - to the third PC, X1v, R1u, R4p+ - to the fourth PC and G3s, R8v+ - to the fifth PC. In both MLR and PCR models Getaway, WHIM and 3D-MoRSE descriptors are found to be important.

Interestingly, the descriptors found to be important in th PCR are different in comparison to the descriptors chosen by the genetic algorithm. The MLR model 1 which includes four descriptors has slightly better statistical results in comparison with the PCR model, with five principal components. Low toxicity values of phosphonium salts are related to the BEHm3, VRD1, VRp1 (Burden eigenvalues and eigenvalue-based indices – descriptors reflecting the

topology of the whole molecule), and in addition to the P2u, H1u, HATS6m, HATSp, R1u, R8v+ descriptors. X1v, Mor05e, G3s, Km, HATS6u, H6v, HATS3v, R3v+, R6e and R4p+ descriptors yielded high toxicity values.

Table 4. Regression coefficients obtained by PCR calculations for the first principal component

Intercept	-10.156
X1v	-0.014
BEHm3	0.452
VRD1	0.004
VRp1	0.004
Mor05e	-0.018
P2u	1.792
G3s	-3.445
Km	-0.607
H1u	0.748
HATS6u	-0.271
HATS6m	1.437
H6v	-2.403
HATS3v	-2.793
HATSp	0.069
R1u	1.375
R3v+	-20.031
R8v+	85.114
R6e	-0.259
R4p+	-11.238

* Dragon descriptors

4. Conclusion

Multiple linear regressions combined with genetic algorithm for variable selection was used to correlate the logarithm of the intravenous lethal dose for mouse with structural features of quaternary phosphonium salts. Following descriptors: topological, Randic molecular profiles geometrical, RDF, 3D-MoRSE, WHIM, Getaway descriptors, functional group counts and atom-centred fragments were present in the final MLR models with acceptable statistical results. The MLR models were then compared to those obtained by principal component regression. The most stable models obtained by both approaches had comparable statistical results. In both MLR and PCR models Getaway, WHIM and 3D-MoRSE descriptors are found to be important for the phosphonium salt toxicity

References

Dizman, B.; Elasri, M.O.; Mathias, L.J. (2004) Synthesis and antimicrobial activities of new water-soluble bis-quaternary ammonium methacrylate polymers. *J. Appl. Polym. Sci.* **94**, 635-642.

Efron, B. (1982) The Jackknife, the Bootstrap and Other Resampling Planes. Society for Industrial and Applied Mathematics, Philadelphia (PA), U.S.A.

Efron, B. (1983) Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Am. Stat. Assoc.* **78**, 316-331.

Fruijtier-Polloth, C. (2005) Safety assessment on polyethylene glycols (PEGs) and their derivatives as used in cosmetic products, *Toxicology* **214**, 1–38.

Gentleman, J.F.; Wilk, M.B. (1975) Detecting outliers. II. Supplementing the direct analysis of residuals. *Biometrics* **31**, 387-410.

Kanazawa, A.; Ikeda, T.; Endo, T. (1993) Novel polycationic biocides: Synthesis and antibacterial activity of polymeric phosphonium salts. *J. Polym. Sci. Part A: Polym. Chem.* **31**, 335-343.

Kanazawa, A.; Ikeda, T.; Endo, T. (1994) Polymeric phosphonium salts as a novel class of cationic biocides. X. Antibacterial activity of filters incorporating phosphonium biocides. *J. Appl. Polym. Sci.* **54**, 1305-1310.

Kenawy, E.-R.; Abdel-Hay, F.I.; El-Shanshoury, A.E.-R.R.; El-Newehy, M.H. (2002) Biologically active polymers. V. Synthesis and antimicrobial activity of modified poly(glycidyl methacrylate-co-2-hydroxyethyl methacrylate) derivatives with quaternary ammonium and phosphonium salts, *J. Polym. Sci. Part A: Polym. Chem.* **40**, 2384-2393.

Lenoir, S.; Pagnoulle, C.; Detrembleur, C.; Galleni, M.; Jérome, R. (2006) New antibacterial cationic surfactants prepared by atom transfer radical polymerization. *J. Polym. Sci. Part A: Polym. Chem.* **44**, 1214-1224.

Lindgren, F.; Hansen, B.; Karcher, W.; Sjöström, M.; Eriksson, L. (1996) Model validation by permutation tests: Applications to variable selection. *J. Chemom.* **10**, 521-532.

Mager, P. P.; Rothe, H. (1990) Obscure phenomena in statistical analysis of quantitative structure-activity relationships. Part 1: Multicollinearity of physicochemical descriptors. *Pharmazie* **45**, 758-764.

Mager, P. P. (1994) Nonleast-squares regression analysis applied to organic and medicinal chemistry. *Med. Res. Rev.* **14**, 533-588.

Popa, A.; Davidescu, C.M.; Ilia, G.; Iliescu, S.; Dehelean, G.; Trif, R.; Păcureanu, L.; Macarie, L. (2003). Synthesis, characterization and antibacterial activity of quaternary phosphonium salts bonded on poly(oxyethylene)s *Rev. Roum. Chim.* **48**(1), 41-48.

Popa, A.; Trif ,A.; Curtui, V.G.; Dehelean, G.; Iliescu, S.; Ilia G. (2002) Synthesis and toxicological characterization of poly(oxyethylene)s functionalized with quaternary phosphonium end groups. *Phosphorus Sulfur*. **177**, 2195-2196.

Schuur, J., Selzer, P.; Gasteiger, J. (1996) The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure–spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci*., **36**, 334–344.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser. B* **36**, 111-147.

Todeschini, R.; Consonni, V.; Maiocchi, A. (1998) The K Correlation index: Theory development and its applications in chemometrics. *Chemometr. Intell. Lab.* **46**, 13-29.

Todeschini, R.; Consonni, V. (2000) *Handbook of Molecular Descriptors,* Wiley, Weinheim, p. 369.

Todeschini, R.; Consonni, V.; Mauri A.; Pavan M. (2003) MobyDigs: software for regression and classification models by genetic algorithms in: *Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Network* (Leardi, R., Ed.), Chapter 5, Elsevier, p. 141-167.

Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. (2004) Detecting 'bad' regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta.* **515**, 199-208.

Wold, S. (1978) Cross validatory estimation of the number of components in factor and principal components models. *Technometrics* **20**, 397-405.

Wold, S.; Dunn III W.J. (1983) Multivariate quantitative structure-activity relationships (QSAR): conditions for their applicability. *J. Chem. Inf. Comput. Sci.* **23**, 6-13.

要　旨

計算科学の手法による有機フォスフォニウム塩の毒性評価

Simona FUNAR-TIMOFEI, Adriana POPA, 鈴木孝弘

ポリマーと結合させたフォスフォニウム塩は、消毒剤、汚れ止めコーティング、繊維の仕上げ剤、水や空気の消毒などに用いられている。その毒性（マウスの静脈注射による致死量の対数値）はＭＬＲによって一連の有機フォスフォニウム塩の化学構造上の特徴と関連があることがわかっている。有機フォスフォニウム塩の立体構造を分子力学法によって最適化し、種々の記述子を得た。化学構造からの記述子と毒性との定量的な関係を、遺伝的アルゴリズムを用いた重回帰分析（ＭＬＲ）を適用してモデル化を試みた。そのＭＬＲモデルを主成分回帰分析によって得られたモデルと比較した。化合物の毒性には、分子の幾何学的な特性と電子分布が影響していることが明らかになった。