A Study of Structure−Carcinogenic Potency Relationship with Artificial Neural Networks. The Using of Descriptors Related to Geometrical and Electronic Structures

Marjan Vračko

National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, P.O. Box 3430, Slovenia

Received May 21, 1997[⊗]

This contribution is an attempt to estimate carcinogenic potency (measured in TD_{50} dose) of molecules using artificial neural networks (ANN) with counterpropagation learning strategy. Three kinds of descriptors have been tested: geometrical structures of molecules, which have been described with 3D coordinates of all atoms, geometrical structures in combination with atomic charges, and energy spectra of occupied orbitals, i.e., the electronic structures. Structures or structures plus atomic charges have been represented with "spectrum-like" representation, which is suitable as input for ANN modelling. A set of 45 benzene derivatives was considered in this study. The models were able to recognize structures of training set, and a weak correlation between descriptors and carcinogenic potency was found.

1. INTRODUCTION

Modeling of the complex multivariate and multiresponse systems is often employed in chemistry. A multiresponse model is a calculation yielding *n*-dimensional output {Y} as a response on *m*-dimensional input {X}. From a mathematical point of view it can be written as Y = M*X, where M is a linear or nonlinear operator describing the model. In the last few years the use of artificial neural networks (ANN) in the field of nonlinear modeling is becoming increasingly intensive.[1] In a presented contribution we applied ANN with counter-propagation (CP) learning strategy.[2] CP ANN represent a possible extension from unsupervised Kohonen learning to supervised learning strategy, and they have been often employed by treating of problems related to chemistry.[3] Since details have been reported elsewhere, just a short introduction is given in Section 2.1.[2−4]

In model studies, as for example in QSAR (Quantitative−Structure Activity Relationship), a molecule is described with a set of descriptors.[5] Some of descriptors are calculated from structural data of molecules[6] and other ones from results of quantum chemical calculations performed on molecules.[7,8] Quantum chemical results are eigenvalues, i.e., molecular orbital energies, and eigenvectors of Hartree−Fock equations or semiempirical Hamiltonian. In QSAR studies from eigenvalues mostly just HOMO and LUMO energies are considered as descriptors. Furthermore, eigenvectors serve for calculation of another descriptors, like charge distribution, i.e., charges on particular atoms, electric moments, or for estimation of log *C*, log *D*, and pK_a values.[9,10] It is to emphasize that in these studies quantum chemical results are used to calculate a small number of quantities (up to a dozen), and a modeling means a searching for good functional (linear or nonlinear) relation between these quantities and investigated property. Alternatively, the molecules can be represented with a large number of data, as, for example, with a numerical representation of different spectra, where the representation vector contains several hundred data. In this way, for example, mass spectra can be used to estimate the class of toxicity.[11] Another example of this type is CoMFA QSAR (Comparative Molecular Field Analysis)[12] or MSA (Molecular shape analysis).[13] It seems that ANN provide an alternative way to find a relationship between multivariate input and investigated property.[3,4]

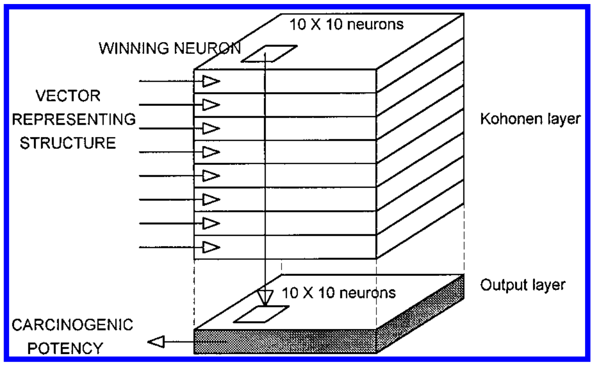

Figure 1. Architecture of CP ANN.

Our goal was to find a representation of geometrical structures and quantum chemical results, which can be directly used as descriptors, i.e., input for modeling. With this intention three kinds of descriptors have been tested, first, spectra related to 3D structures, second, spectra related to 3D structures in combination with atomic charges, and third, spectra of molecular orbital energies (electronic structures). First and second types of representations are described in Section 2.2. The first type describes geometrical and the third one pure quantum chemical features of molecules. The second one is a combination of both; calculated atomic charges are included in a representation of 3D structures.

In this study 45 benzene derivatives have been considered. Investigated property has been the carcinogenic potency[14] measured in logarithm of dose TD_{50} (see Section 2.3).

2. METHODS AND COMPUTATIONAL DETAILS

2.1. Artificial Neural Networks with Counterpropagation Learning Strategy. Architecture of the CP ANN is shown in Figure 1. CP ANN is built up from two layers of neurons arranged in two-dimensional rectangular matrices. The input or Kohonen layer gets input variables related to considered objects. The target values (in our case carcino-

[⊗] Abstract published in *Advance ACS Abstracts,* November 1, 1997.

S0095-2338(97)00231-X CCC: $14.00

Table 1. CAS Numbers and Chemical Names of 45 Compounds

	CAS	name
1	100-00-5	1-chloro-4-nitrobenzene
2	100-51-6	benzyl alcohol
3	101-90-6	diglycidyl resorcinol ether, technical grade
4	102-50-1	*m*-cresidine
5	103-03-7	1-carbamyl-2-phenylhydrazine
6	103-72-0	phenyl isothiocyanate
7	103-85-5	1-phenyl-2-thiourea
8	105-11-3	*p*-quinone dioxime
9	106-46-7	1,4-dichlorobenzene
10	106-47-8	*p*-chloroaniline
11	108-90-7	chlorobenzene
12	118-74-1	hexachlorobenzene
13	118-75-2	chloranil
14	120-71-8	*p*-cresidine
15	123-31-9	hydroquinone
16	135-20-6	cupferron
17	137-17-7	2,4,5-trimethylaniline
18	142-04-1	aniline•HCl
19	5307-14-2	2-nitro-*p*-phenylenediamine
20	59-88-1	phenylhydrazine•HCl
21	613-94-5	benzoyl hydrazine
22	619-17-0	4-nitroanthranilic acid
23	622-51-5	*p*-tolylurea
24	624-18-0	*p*-phenylenediamine•2HCl
25	6334-11-8	2,4,6-trimethylaniline•HCl
26	634-93-5	2,4,6-trichloroaniline
27	6379-46-0	1,2,3-trichloro-4,6-dinitrobenzene
28	6965-71-5	α-(2,5-dichlorophenoxy)propionic acid
29	71-43-2	benzene
30	82-68-8	pentachloronitrobenzene
31	87-86-5	2,3,4,5,6-pentachlorophenol (dowicide ec-7)
32	88-06-2	2,4,6-trichlorophenol
33	88-73-3	1-chloro-2-nitrobenzene
34	93-76-5	2,4,5-trichlorophenoxyacetic acid
35	94-75-7	2,4-dichlorophenoxyacetic acid
36	95-50-1	1,2-dichlorobenzene
37	95-74-9	3-chloro-*p*-toluidine
38	95-79-4	5-chloro-*o*-toluidine
39	95-83-0	4-chloro-*o*-phenylenediamine
40	97-00-7	1-chloro-2,4-dinitrobenzene
41	99-55-8	5-nitro-*o*-toluidine
42	99-57-0	2-amino-4-nitrophenol
43	99-59-2	5-nitro-*o*-anisidine
44	305-03-3	chlorambucil
45	366-70-1	procarbazine•HCl

Table 2. Experimental, Retrieved (Recall Ability), and Predicted (Cross Validation) LgTD_{50} Values Using Structures (S), Structures + Charges (SC), or Electronic Structures as (ES) Descriptor

			S		SC		ES	
	CAS	exp.	ret.	pred.	ret.	pred.	ret.	pred.
1	100-00-5	2.560	2.560	1.946	2.844	2.675	2.560	2.875
2	100-51-6	1.820	2.515	1.018	1.412	1.018	1.820	2.696
3	101-90-6	4.090	4.090	2.833	4.090	2.813	4.090	5.522
4	102-50-1	2.820	2.820	2.177	2.820	2.823	2.820	3.437
5	103-03-7	3.000	3.000	3.059	3.000	2.974	3.000	3.088
6	1 03-72-0	2.990	1.989	2.609	2.990	3.107	2.864	2.699
7	103-85-5	3.060	3.060	2.989	3.060	2.974	3.060	2.506
8	105-11-3	2.170	2.170	3.219	2.170	3.219	2.170	3.194
9	106-46-7	2.640	2.995	3.103	2.995	3.103	2.500	2.363
10	106-47-8	1.940	1.900	1.861	1.900	1.861	1.940	2.228
11	108-90-7	3.270	2.995	3.179	2.995	2.904	3.270	3.000
12	118-74-1	3.790	3.790	3.478	3.790	3.110	3.790	4.176
13	118-75-2	3.630	3.630	3.245	3.630	3.265	3.630	3.542
14	120-71-8	3.490	3.490	2.213	3.490	2.426	3.490	2.877
15	123-31-9	3.230	3.230	2.748	3.230	2.956	3.230	2.696
16	135-20-6	2.740	2.740	2.517	2.740	2.069	2.864	3.091
17	137-17-7	4.340	4.340	2.869	4.340	2.863	4.340	3.254
18	142-04-1	1.010	1.989	2.975	1.412	1.812	1.010	3.451
19	5307-14-2	2.400	2.400	1.856	2.400	1.856	2.400	2.657
20	59-88-1	3.200	2.515	2.294	3.200	2.126	3.458	2.462
21	613-94-5	4.270	4.270	2.527	4.270	2.986	4.270	3.133
22	619-17-0	1.160	1.160	2.160	1.160	2.784	1.160	2.611
23	622-51-5	2.860	2.860	1.909	2.860	3.197	2.860	2.630
24	624-18-0	1.860	1.900	1.939	1.900	1.939	1.860	2.678
25	6334-11-8	3.870	3.870	3.365	3.870	2.908	3.870	3.938
26	634-93-5	2.880	3.174	3.454	2.880	3.454	2.880	2.977
27	6379-46-0	3.740	3.740	3.621	3.740	3.621	3.740	3.370
28	6965-71-5	3.810	3.810	3.256	3.810	3.256	3.810	4.757
29	71-43-2	3.710	2.995	2.767	2.995	2.767	3.458	2.020
30	82-68-8	3.620	3.620	3.436	3.620	3.383	3.620	3.675
31	87-86-5	4.180	4.180	3.713	4.180	3.794	4.180	4.086
32	88-06-2	3.460	3.174	2.886	3.460	3.312	3.460	3.387
33	88-73-3	3.160	3.160	3.212	3.160	3.056	3.160	2.566
34	93-76-5	4.350	4.350	2.658	4.350	3.349	4.350	3.261
35	94-75-7	3.250	3.250	2.781	3.250	3.145	3.250	4.339
36	95-50-1	2.360	2.995	2.986	2.995	3.149	2.500	2.637
37	95-74-9	1.850	1.850	2.395	1.850	2.395	1.850	2.647
38	95-79-4	3.020	2.908	2.802	2.908	2.802	3.020	2.463
39	95-83-0	2.170	2.170	2.473	2.170	3.011	2.170	2.627
40	97-00-7	3.130	2.703	2.298	2.844	2.566	3.130	3.734
41	99-55-8	2.800	2.908	3.018	2.908	3.018	2.800	2.447
42	99-57-0	2.290	2.703	3.122	2.244	2.201	2.290	2.201
43	99-59-2	2.200	2.200	2.897	2.244	2.289	2.200	2.289
44	305-03-3	6.500	6.500	6.064	6.500	6.064	6.500	5.141
45	366-70-1	6.060	6.060	4.766	6.060	3.625	6.060	4.900

genic potency) are during the learning given to output layer, which has the same topological arrangement of neurons as Kohonen layer. Learning in Kohonen layer is the same as in Kohonen networks. This means a vector of input variables is presented to all neurons. Program selects the neuron, which weights are closest to the input values. The chosen neuron is called the winning neuron. The position of the winning neuron is transferred from the Kohonen layer to the output layer, and the weights in output layer are corrected according to the given target values. Step by step the weights in Kohonen and output layer are corrected in such a way that they are becoming similar to the input values. After the weights are stabilized the CP ANN is considered to be trained. In the prediction phase the carcinogenic potency values are taken from the output layer. *Hecht-Nielsen*[15] and *Dayhof*[16] give a detailed description of CP ANN architecture and learning strategy. In our case, if a new structure with unknown carcinogenic potency is presented to the system, it will be first situated into the Kohonen layer, the found position will be projected to the output layer, and from this position in the output layer the predicted value will be drawn. In other words, the ability of CP ANN is a clustering of compounds respecting input variables. Compounds with similar representation vectors are situated close to each other in a two-dimensional network.

2.2. "Spectrum-like" Representation of Molecular Structure.[17] In computational treatment of molecular structures the problem of representation plays a central role. Molecules must be represented in such a way that computer code can recognize similarities and find common patterns in different structures. On the other hand, the representation must satisfy the following requirements:

−uniqueness (different code for different structures),

−number, type and domain of variables are the same for all structures,

−reversibility (by proper orientation the structure can be obtained back from the code).

For example, standard description of geometrical structures where the positions of atoms are given in Cartesian coordinates does not satisfy the second requirement. Namely, dimension of representation depends on the number of atoms

Table 3. Set of 45 Compounds: CAS Numbers of Compounds That Cannot Be Distinguished by Models[a]

R	S CAS	R	SC CAS	R	ES CAS
2.515	100-51-6 (1.820)	2.844	100-00-5 (2.560)	2.864	103-72-0 (2.990)
	59-88-1 (3.200)		97-00-7 (3.130)		135-20-6 (2.740)
1.989	103-72-0 (2.990)	1.412	100-51-6 (1.820)	2.500	106-46-7 (2.640)
	142-04-1 (1.010)		142-04-1 (1.010)		95-50-1 (2.360)
2.995	106-46-7 (2.640)	2.995	106-46-7 (2.640)	3.458	59-88-1 (3.200)
	108-90-7 (3.270)		108-90-7 (3.270)		71-43-2 (3.710)
	71-43-2 (3.700)		71-43-2 (3.700)		
	95-50-1 (2.360)		95-50-1 (2.360)		
1.900	106-47-8 (1.940)	1.900	106-47-8 (1.940)		
	624-18-0 (1.860)		624-18-0 (1.860)		
3.174	634-93-5 (2.880)	2.244	99-57-0 (2.290)		
	88-06-2 (3.460)		99-59-2 (2.200)		
2.703	97-00-7 (3.130)	2.908	95-79-4 (3.020)		
	99-57-0 (2.290)		99-55-8 (2.800)		

[a] Experimental LgTD_{50} values in brackets, R retrieved values, S structures, SC structures plus atomic charges, ES electronic structures as descriptors.

in molecule. It can be shown that "spectrum-like" representation is the suitable one.

The "spectrum-like" representation of molecules is obtained by projection of a molecule onto a sphere of arbitrary radius. First, an oriented structure is situated into the sphere. The projection beam from the central point of the sphere (gnomonic projection) causes a pattern of points on the sphere, where each point represents the particular atom. Second, each point is taken as the center of a "bell-shaped" function with intensity related to the distance between the coordinate origin and a particular atom. As additional parameter atomic charge can be incorporated into this function. As a "bell-shaped" function we have taken the Lorentzian function with the form

$$s_i(\varphi_j,\theta_l) = \frac{\rho_i}{(\varphi_j - \varphi_i)^2 + \sigma_i^2} + \frac{\rho_i}{(\theta_l - \theta_i)^2 + \sigma_i^2}$$

$$j = 0,\frac{2\pi}{k},...\frac{2\pi}{k}j,....2\pi; l = 0,...\frac{\pi}{k/2}l,...\pi \quad (1)$$

Here is $s_i(\varphi_j,\theta_l)$, intensity related to atom i; ρ_i, distance between zero point and atom i; σ_i, atomic charge + 1 on atom i; θ_i,φ_i, azimuthal and polar angle of atom i; k, an integer number (see text below). σ_is are set to one if only atom positions are considered. The total intensity related to the entire molecule is the sum of densities belonging to individual atoms. This representation fulfills all demands mentioned above.

Even if atomic charges are incorporated in eq 1, the reversibility is not lost; however, recovering of atom positions from the code is not straightforward any more and must be done in an iterative way.

Unfortunately, this representation is not invariant on rotation. This fact requires proper orientation of all molecules in the study. For molecules with a common substructure it is convenient to select an orientation rule to superposition of substructures.

2.3. Databases. Geometrical structures of investigated molecules have been taken from NCI DIS 3D 127k database (National Cancer Institute Drug Information System).[18] The NCI DIS database is a collection of 2D structure representations for over 400 000 drugs. The structural information stored in NCI DIS is only the connection table for each drug. The NCI DIS 3D 127k database contains about 127 000 3D molecular structures. Program Corina[19] was used to convert the connection tables into 3D structures.

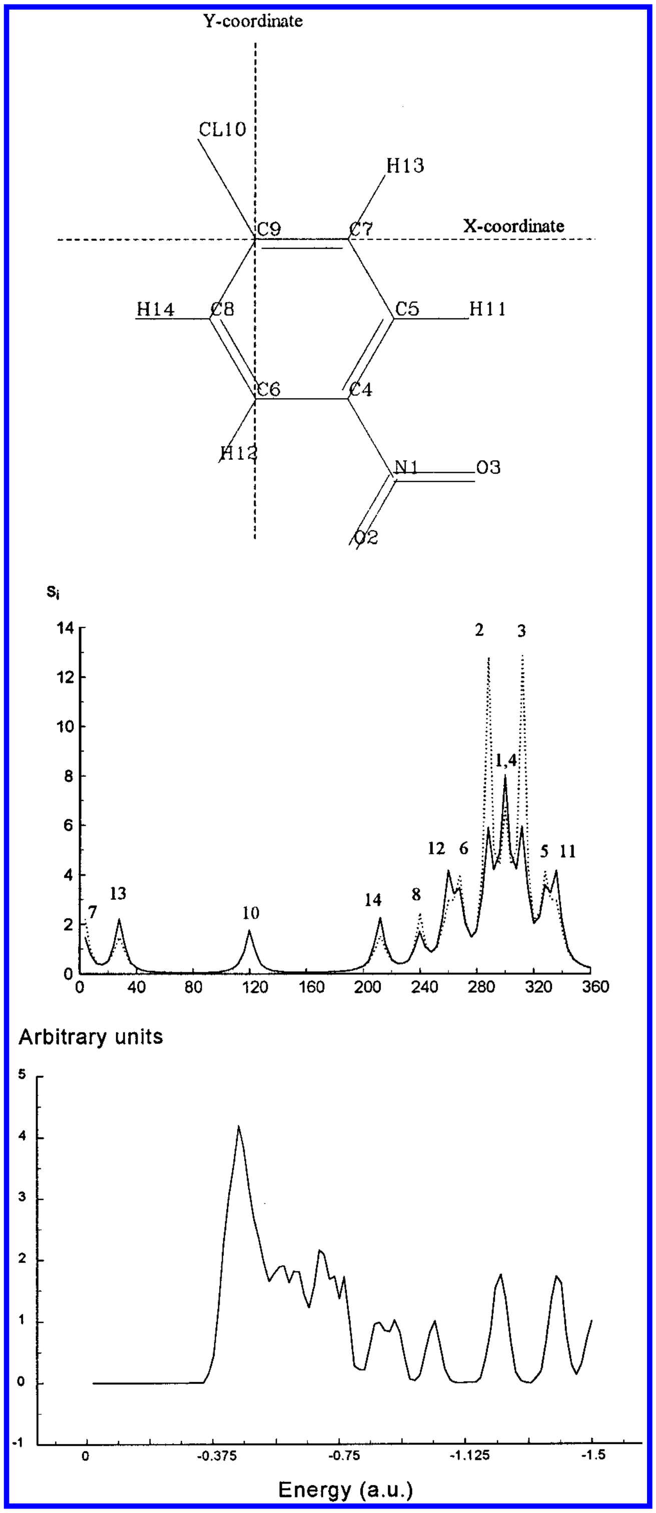

Figure 2. Representations of 1-chloro-4-nitrobenzene: A: Orientation of structure, zero point is set on carbon atom with chlorine atom. B: "Spectrum-like" representation of structure (full line) and structure + atomic charges (dot line), numbers 1−14 denote atoms. C: Electronic structure.

Carcinogenic potency data have been taken from the Gold database, which contains names, CAS registration numbers, and carcinogenic potency data for about 1100 compounds.[20] Carcinogenic potency data are given for diverse animals and tissues. An upgraded version of the Gold database[21] contains uniform parameters of carcinogenic potency for all compounds measured as dose TD_{50} for mice or rats. TD_{50} is

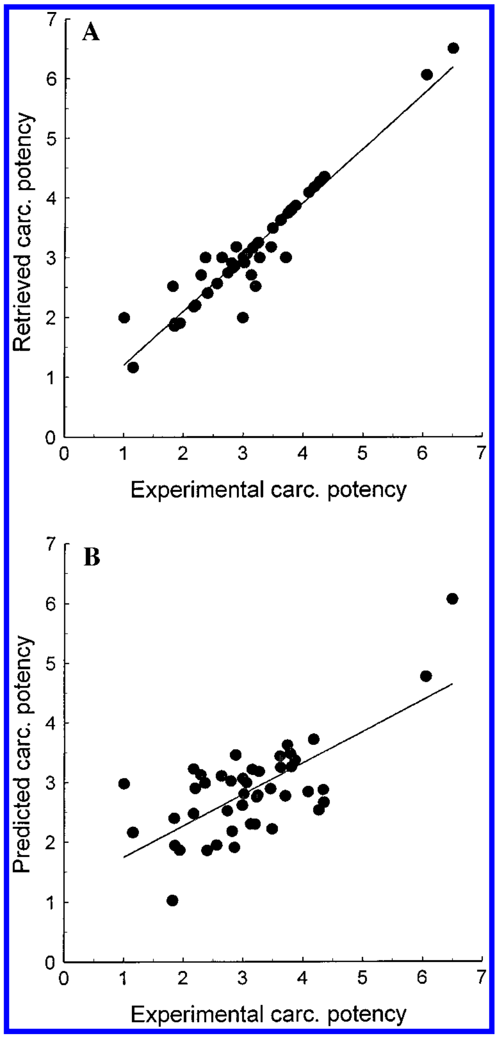

Figure 3. Predicted versus experimental LgTD$_{50}$ values for a set of 45 compounds with structures as descriptors. A: values retrieved from model ($r = 0.95$, $b_0 = 0.29$, $b_1 = 0.91$) and B: cross validation results ($r = 0.70$, $b_0 = 1.21$, $b_1 = 0.53$).

given in mg per kg of animal weight,[20,22] but in our studies the units have been transformed to logarithmic values LgTD$_{50}$ = log (MW*1000/TD$_{50}$) where MW is molecular weight. This unit is more related to toxicity or carcinogenic potency.[23] LgTD$_{50}$ for mice has been taken as investigated property in this study.

From the Gold database we selected a set of 45 benzene analogues, whose CAS registration numbers and names are given in Table 1. Compounds from the Gold database have been often treated by theoretical methods. *Gombar et al.*[24] studied a set of 269 compounds belonging to different chemical groups by QSAR method. Our set includes 25 compounds from this set; another 20 are taken from the upgraded Gold database.

2.4. Some Computational Details. All compounds in our study contain a benzene ring as a common substructure. The orientation rule was chosen for the superposition of benzene rings, which were situated into an $x-y$ plane. Coordinate zero point was set on the carbon atom of benzene ring that carries a substituent with highest priority under IUPAC (International Union of Pure and Applied Chemistry) nomenclature. Because selected molecules are planar or close-to-planar their structures are sufficiently described only by $x-y$ coordinates. In this case, rather than the projection on an entire sphere just the projection on an equatorial circle has been taken as a representative spectrum. By selection of k equidistant points on an interval $(0, 2\pi)$ the structure is represented with k-dimensional vector. (It is obvious that k must be larger than the number of atoms in the largest molecule.) In this study k has been set to a value of 90.

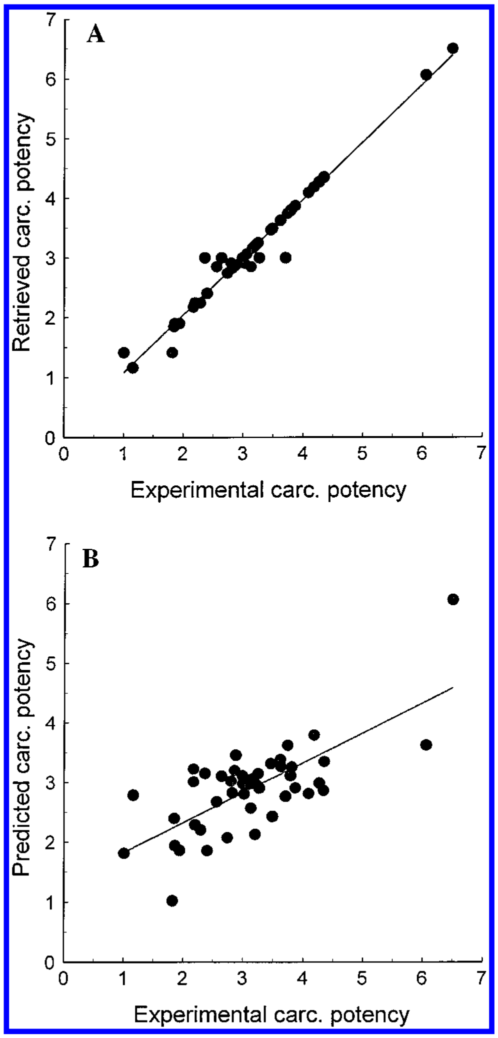

Figure 4. Predicted versus experimental LgTD$_{50}$ values for a set of 45 compounds with structures plus atomic charges as descriptors: A: values retrieved from model ($r = 0.98$, $b_0 = 0.10$, $b_1 = 0.97$) and B: cross validation results ($r = 0.72$, $b_0 = 1.32$, $b_1 = 0.50$).

Atomic charges and orbital energies (spectra of occupied orbitals) have been calculated in AM1 approximation using the GAUSSIAN 94 program package.[25] Spectra of ES have been obtained by the folding of discrete lines corresponding to orbital energies with Gaussian functions. Eighty points were taken on an energy scale between -1.5 au (atomic unit) to 0 au.

A network was built from 100 neurons arranged in a 10 × 10 rectangular. Models were obtained by training of networks with 400 learning epochs.

Models were tested on recall ability and with cross validation method. The recall ability of a model is defined as the ability of a model to recognize the objects from a training set. It provides information on how well the model fits to training data but no information about prediction ability. With a cross validation method the robustness of a model is tested. In this procedure the objects one after another are selected out, whereas for every selected object

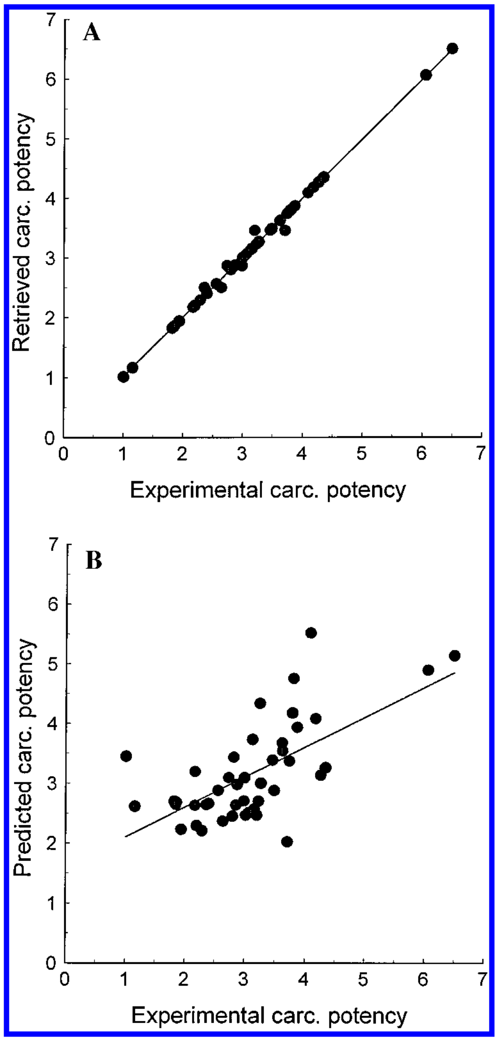

Figure 5. Predicted versus experimental LgTD$_{50}$ values for a set of 45 compounds with electronic structure spectra as descriptors: A: values retrieved from (r = 1.00, b_0 = 0.01, b_1 = 1.00) and B: cross validation results (r = 0.63, b_0 = 1.59, b_1 = 0.50).

the model is built up with remaining ones. Furthermore, this model is used to predict the value for selected object. It gives us information on the prediction ability and on the quality of selected training set, i.e., it shows the outliers. As statistical parameters we report the correlation coefficient r and the parameters of line (b_0, b_1) showing predicted versus experimental values.[26] In an ideal case, i.e., all experimental and predicted values are the same, $r = 1$, $b_0 = 0$, $b_1 = 1$.

3. RESULTS AND DISCUSSION

Results for a set with 45 compounds are shown in Tables 2 and 3 and in Figures 3A–5A. Figures 3A–5A show recall ability of models with different descriptors. Table 3 shows the compounds that cannot be distinguished by models. Such compounds have the same retrieved value. The inability to distinguish the compounds of a training set is not necessarily a shortcoming of a model. It simply means that the descriptors of some compounds are too similar to be discriminated by the model. Generally, the statistical parameters of recall ability tests are good for an entire set ($r > 0.9$, $b_0 \cong 0.1$, $b_1 > 0.9$). Cross validation predicted values are shown in Table 2 and Figures 3B–5B. Statistical parameters are $r \cong 0.7$, $b_0 \cong 1.5$, and $b_1 \cong 0.5$.

Table 4. Experimental, Retrieved (Recall Ability), and Predicted (Cross Validation) LgTD$_{50}$ Values Using Structures (S), Structures + Charges (SC), or Electronic Structures as (ES) Descriptor[a]

			S		SC		ES	
	CAS	exp.	ret.	pred.	ret.	pred.	ret.	pred.
1	100-00-5	2.560	2.560	2.292	2.844	3.124	2.860	2.611
2	100-51-6	1.820	*	*	*	*	*	*
3	101-90-6	4.090	4.090	2.833	4.090	2.813	4.090	4.932
4	102-50-1	2.820	2.820	2.177	2.820	2.553	2.820	3.177
5	103-03-7	3.000	3.000	3.059	3.000	3.198	3.000	3.045
6	103-72-0	2.990	2.990	3.211	2.990	2.995	2.864	2.723
7	103-85-5	3.060	3.060	3.199	3.060	3.199	3.060	3.158
8	105-11-3	2.170	2.170	3.219	2.170	3.507	2.170	2.855
9	106-46-7	2.640	2.995	3.103	2.995	3.103	2.500	2.363
10	106-47-8	1.940	1.940	2.205	1.940	1.861	1.940	2.168
11	108-90-7	3.270	2.995	2.904	2.995	2.904	3.270	2.418
12	118-74-1	3.790	3.709	3.830	3.790	3.244	3.790	4.176
13	118-75-2	3.630	3.709	3.241	3.630	3.335	3.630	3.107
14	120-71-8	3.490	3.490	2.213	3.490	2.476	3.490	3.252
15	123-31-9	3.230	3.230	2.412	3.709	4.170	3.230	3.792
16	135-20-6	2.740	2.740	3.195	2.740	2.641	2.864	3.147
17	137-17-7	4.340	4.340	2.868	4.340	3.102	4.340	3.316
18	142-04-1	1.010	*	*	*	*	*	*
19	5307-14-2	2.400	2.400	2.289	2.400	1.856	2.400	2.524
20	59-88-1	3.200	3.200	2.745	3.200	3.553	3.200	1.873
21	613-94-5	4.270	4.270	2.986	4.270	2.755	4.270	2.995
22	619-17-0	1.160	*	*	*	*	*	*
23	622-51-5	2.860	2.860	1.909	2.860	3.197	2.860	2.917
24	624-18-0	1.860	1.860	2.242	1.860	1.939	1.860	2.731
25	6334-11-8	3.870	3.870	3.301	3.870	4.335	3.870	2.696
26	634-93-5	2.880	3.174	3.454	2.880	3.454	2.880	2.613
27	6379-46-0	3.740	3.740	3.621	3.740	3.621	3.740	3.445
28	6965-71-5	3.810	3.810	3.256	3.810	3.256	3.810	3.467
29	71-43-2	3.710	2.995	2.767	2.995	2.767	*	*
30	82-68-8	3.620	3.620	3.788	3.620	3.271	3.620	3.585
31	87-86-5	4.180	4.180	3.713	3.709	3.794	4.180	3.622
32	88-06-2	3.460	3.174	3.026	3.460	3.356	3.460	3.242
33	88-73-3	3.160	3.160	3.049	3.160	2.996	2.860	2.566
34	93-76-5	4.350	4.350	2.658	4.350	3.577	4.350	3.261
35	94-75-7	3.250	3.250	4.052	3.250	3.145	3.250	4.339
36	95-50-1	2.360	2.995	3.204	2.995	3.149	2.500	2.637
37	95-74-9	1.850	1.850	2.395	1.850	2.395	1.850	2.378
38	95-79-4	3.020	3.020	2.802	3.020	2.508	3.020	2.315
39	95-83-0	2.170	2.170	2.918	2.170	2.918	2.170	2.642
40	97-00-7	3.130	3.130	2.425	2.844	2.566	3.130	3.734
41	99-55-8	2.800	2.800	3.018	2.800	3.018	2.800	2.657
42	99-57-0	2.290	2.290	3.122	2.244	2.201	2.290	2.201
43	99-59-2	2.200	2.200	3.477	2.244	2.289	2.200	2.498
44	305-03-3	6.500	6.500	6.064	6.500	6.064	6.500	5.141
45	366-70-1	6.060	6.060	6.496	6.060	4.920	6.060	4.591

[a] *denotes outliers.

Table 5. Set of 42 (41) Compounds: CAS Numbers of Compounds That Cannot Be Distinguished by Models[a]

R	S CAS	R	SC CAS	R	ES CAS
2.995	106-46-7 (2.640)	2.844	100-00-5 (2.560)	2.860	100-00-5 (2.560)
	108-90-7 (3.270)		97-00-7 (3.130)		622-51-5 (2.860)
	71-43-2 (3.700)	2.995	106-46-7 (2.640)		88-73-3 (3.160)
	95-50-1 (2.360)		108-90-7 (3.270)	2.864	103-72-0 (2.990)
3.709	118-74-1 (3.790)		71-43-2 (3.700)		135-20-6 (2.740)
	118-75-2 (3.630)		95-50-1 (2.360)	2.500	106-46-7 (2.640)
3.174	634-93-5 (2.880)	3.709	123-31-9 (3.230)		95-50-1 (2.360)
	88-06-2 (3.460)		87-86-5 (4.180		
		2.244	99-57-0		
			99-59-2		

[a] Experimental LgTD$_{50}$ values in brackets.

In the next step some of the compounds were selected out as outliers. Compounds were not removed in order to achieve better statistical parameters, but they are excluded because their toxic character cannot be described with other compounds in the set. The compounds are (with experi-

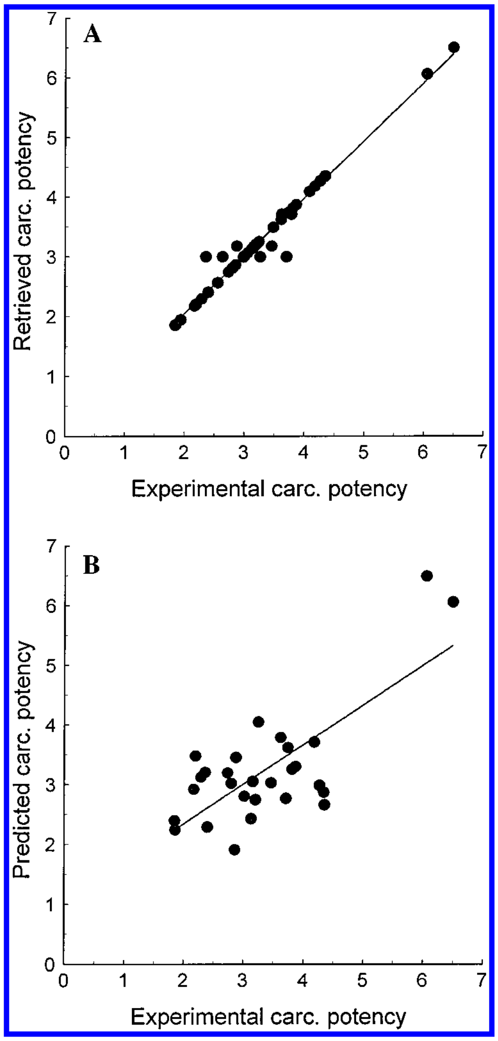

Figure 6. Predicted versus experimental LgTD$_{50}$ values for a set of 42 compounds of set 2 with structures as descriptors: A: values retrieved from model (r = 0.98, b_0 = 0. 11, b_1 = 0.97) and B: cross validation results (r = 0.74, b_0 = 1.01, b_1 = 0.66).

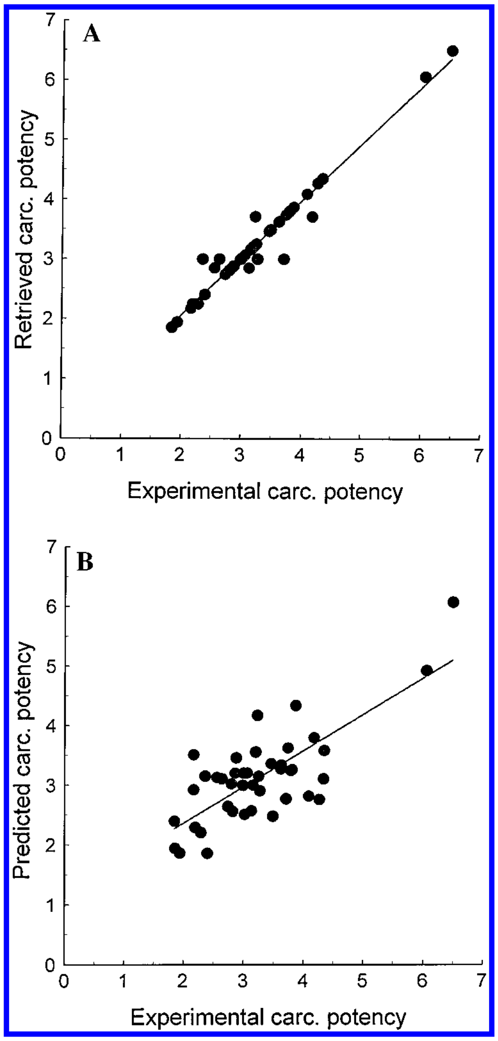

Figure 7. Predicted versus experimental LgTD$_{50}$ values for a set of 42 compounds with structures plus charges as descriptors: A: values retrieved from model (r = 0.98, b_0 = 0.15, b_1 = 0.96) and B: cross validation results (r = 0.76, b_0 = 1.14, b_1 = 0.61).

mental LgTD$_{50}$ values in brackets) benzyl alcohol (1.820), aniline•HCl (1.010), benzoyl hydrazine (1.160), and, by considering of electronic structures as descriptors, additionally α-(2,5-dichlorophenoxy)propionic acid (3.710). Recall ability is enhanced a little bit ($r > 0.98$, $b_0 \cong 0.1$, $b_1 > 0.96$), but still some of the compounds cannot be distinguish by models (see Table 5). Enhanced is also the prediction ability of models with statistical parameters: $r \cong 0.74$, $b_0 \cong 1.2$, $b_1 \cong 0.6$ (see Table 4 and Figures 6B−8B).

4. CONCLUSIONS

It is a difficult task to find a good correlation between molecular structures and actual carcinogenic potency. The term “carcinogenic potency” defines just a biological endpoint, which indeed includes several different mechanisms. Experimental evaluation of carcinogenic potency is difficult because several different conditions, like sex, age, and health of testing animals, influence the results. Therefore, the reliability of experimental data is for some compounds questionable. On the other hand, this topic is quite actual because of the high costs and long duration of animal experiments as well as of ethical reasons, which demand reduction (or complete abolishing) of experiments on mammals.

In the present study a set of 45 benzene analogues was treated with CP ANN method to build models for prediction of carcinogenic potency. Molecules have been described with their geometrical structures, or with geometrical structures in combination with atomic charges, or with electronic structure spectra. These descriptors carry the entire information about structural and electronic properties of molecules, which determine molecular behavior in a biochemical environment.[27] Two questions have been addressed: first, can CP ANN recognize the compounds from a training set, and, second, are the models suitable for prediction of carcinogenic potency. The answer for the first question is that compounds described with mentioned descriptors can be distinguished by CP ANN. The best recall ability was found with electronic structure spectra as descriptors. Second, the prediction ability of models was too poor to get reliable numbers for a carcinogenic potency dose. But a general trend is evident, i.e., models predict higher carcinogenicity for compounds with higher experimental values. On the other hand, any conclusions about mechanisms of carcinogenicity would be too vague at this point in time.

A crucial point in building of models is selection of compounds, i.e., the selection of training set. To get more reliable results more compounds must be included into a training set. This work is in progress in our laboratory.

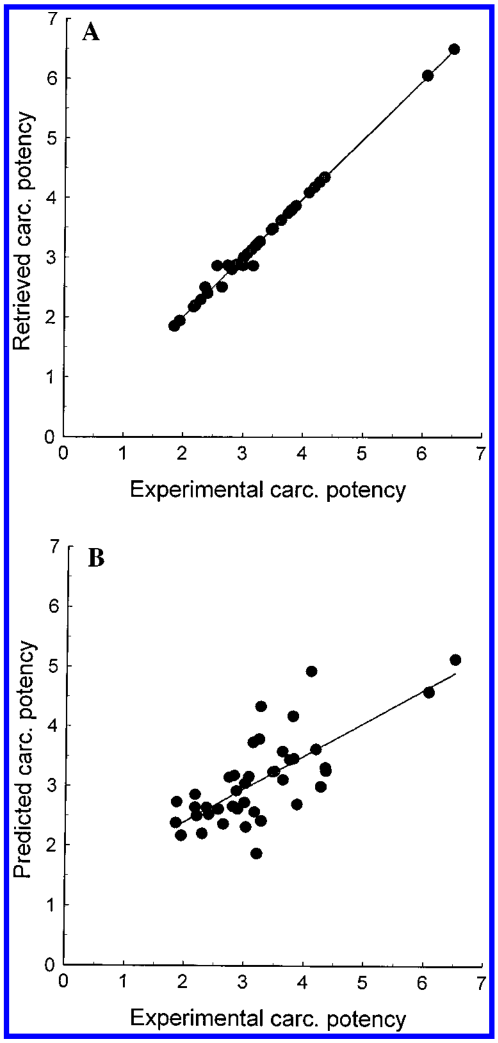

Figure 8. Predicted versus experimental $LgTD_{50}$ values for set of 41 compounds with electronic structure spectra as descriptors: A: values retrieved from ($r = 1.0$, $b_0 = 0.02$, $b_1 = 1.0$) and B: cross validation results ($r = 0.73$, $b_0 = 1.29$, $b_1 = 0.56$).

ACKNOWLEDGMENT

The author gratefully thanks European Union that supports this work under contract COPERNICUS EST CP94 1029 and also thanks Ministry of Science and Technology of the Republic of Slovenia obtained by the Project J1-5014-104 for additional support.

REFERENCES AND NOTES

(1) Villemin, D.; Cherqaoui, D.; Mesbah, A. Predicting Carcinogenicity of Polycyclic Aromatic Hydrocarbons from Back-Propagation Neural Network. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1288−1293.

(2) Zupan, J.; Novič, M.; Gasteiger, J. Neural networks with counter-propagation learning strategy used for modelling. *Chemom. Intell. Lab. Syst.* **1995**, *27*, 175−187.

(3) Novič, M.; Zupan, J. Investigation of Infrared Spectra-Structure Correlation Using Kohonen and Counterpropagation Neural Network. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 454−466.

(4) Zupan, J.; Gasteiger, J. *Neural networks for chemists: an introduction*; VCH Publishers: 1992; p 99.

(5) Hansch, C.; Leo, A. Exploring QSAR. Fundamentals and Applications in Chemistry and Biology; ACS Professional Reference Book, American Chemical Society: Washington, DC, 1995; p l9.

(6) Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Chemical Information in 3D Space. J. *Chem. Inf. Comput. Sci.* **1996**, *36*, 1030−1037.

(7) Karelson, M.; Lobanov, V. S. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027−1043.

(8) Bruno-Blanch, L.; Estiu, G. L. Quantum Chemistry in QSAR: Anticonvulsivant Activity of VPA Derivatives. *Int. J. Quantum Chem: Quantum Biology Symp.* **1995**, *22*, 39−49.

(9) Klopman, G.; Iroff, L. D. Calculation of Partition Coefficients by the Charge Density Method. *J. Comput. Chem.* **1981**, *2*, 157−160.

(10) Houser, J. J.; Klopman, G. A New Tool for the Rapid Estimation of Charge Distribution. *J. Comput. Chem.* **1988**, *8*, 893−904.

(11) Yin, L.-B. Classification results of Mass Spectra of Toxic Compounds by Class Modeling. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1232−1234.

(12) Cramer, R. D. III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(13) Rhyu, K. B.; Patel, H. C.; Hopfinger, A. J. A 3D-QSAR Study of Anticoccidial Triazines Using Molecular Shape Analysis. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 771−778.

(14) Benigni, R.; Giuliani, A. Quantitative Structure-Activity relationship (QSAR) Studies of Mutagens and Carcinogens. *Med. Res Rev.* **1996**, *16*, 267−284.

(15) Hecht-Nielsen, R. Counter propagation Networks. *Appl. Optics* **1987**, *26*, 4979−4984.

(16) Dayhof, J. *Neural Network Architectures, An Introduction*; Van Nostrand Reinhold: New York, 1990; p 192.

(17) Novič, M.; Zupan, J. A New General and Uniform Structure Representation. In *Information und Wissen am Arbeitsplatz des Chemikers*; Jahrestagung 1995, CIC-Workshop, Hochfilzen/Tirol, November 1995, 19−21.

(18) Milne, G. W. A.; Nicklaus, M. C.; Driscoll, J. S.; Wang, S.; Zaharevitz, D. National Cancer Institute Drug Information System 3D Database. J. *Chem. Inf. Comput. Sci.* **1994**, *34*, 1219−1224.

(19) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(20) Gold. L. S.; Sawyer, C. B.; Magaw, R.; Backman, G. M.; de Veciana, M.; Levinson, R.; Hooper, N. K.; Havender, W. R.; Bernstein, L.; Peto, R.; Pike, M. C.; Ames, B. N. A Carcinogenic Potency Database of the standarized results of animal bioassays. *Environ. Health Persp.* **1986**, *58*, 9−319.

(21) Gold, L. S.; Manley, N. B.; Slone, T. H.; Garfinkel, G. B.; Ames, B. N.; Rohrbach, L.; Stern, B. R.; Chow, K. Sixth Plot of the Carcinogenic potency database: Reults of Animal Bioassays published in the general Literature 1989 to 1990 and by the National Toxicology Program 1990 to 1993. *Environ. Health Persp.* **1995**, *103*, Suppl. 8, 3−122.

(22) Peto, R.; Pike, M. C.; Bernstein, L.; Gold, L. S.; Ames, B. N. The TD_{50}: a proposed general convention for the numerical description of the carcinogenic potency of chemicals in chronic-exposure animal experiments. *Environ. Health Persp.* **1984**, *58*, 1−8.

(23) Kamlet, M. J.; Doherty, R. M.; Taft, R. W.; Abraham, M. H.; Veith, G. D.; Abraham, D. J. Solubility Properties in Polymers and Biological Media. 8. An Analysis of the Factors that Influence Toxicities of organic Nonelectrolytes to the Golden Orfe Fish (Leuciscus idus melanotus). *Environ. Sci. Technol.* **1987**, *21*, 149−155.

(24) Gombar, V. K.; Enslein, K.; Hart, J. B.; Blake, W.; Borgstedt, H. H. Estimation of Maximum Tolerated Dose for Long-Term Bioassays from Acute Lethal Dose and Structure by QSAR. *Risk Analysis* **1991**, *11*, 509−517.

(25) GAUSSIAN 94, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon M.; Gonzalez, C.; Pople, J. A. Gaussian Inc.: Pittsburgh, PA, 1995.

(26) Mendenhall, W.; Sincich, T. Statistics for Engineering and the Sciences; Prentice-Hall International, Inc.: 1995; pp 279, 531.

(27) Gough, K. M.; Belohorcova, K.; Kaiser, K. L. E. Quantitative Structure-Activity Relationship (QSARs) of Photobacterium phosphoreum toxicity of nitrobenzene derivatives. *Sci. Total Environ.* **1994**, *142*, 179−190.

CI970231Y