Application of Artificial Neural Networks to the QSPR Study – Automated Classification of Endocrine Disrupting Chemicals

KUI 14/2004
Received December 12, 2003
Accepted May 18, 2004

*M. Novič and A. Roncaglioni**

National Institute of Chemistry, Ljubljana, Slovenia
*Institute Mario Negri, Milano, Italy

The European Union Commission delivered a list of 553 chemicals that were inspected for the scientific evidence of their endocrine disruption activity. The source of information, i.e. the studies collected in the report refer to experiments made during the decades, evaluating several species and a great variety of effects, which reflects in non-homogeneity of the data. The classification of potential endocrine disrupters (EDs), according to the literature evidence of their functioning, was proposed by the Commission. The endocrine disruption categories given in the EU Commission report are the following: (i) certainly active as endocrine disrupters, (ii) potentially active, (iii) less probable active – lacking evidence, and (iv) certainty non-active. The research of the methodology to find an automated predictive model, yielding the ED categories, is presented. Clustering and classification techniques were employed to solve the problem. From the list of 553 chemicals a dataset of 106 molecules with the defined chemical structure and ED class were extracted. Molecular structures were represented by 3D atomic coordinates calculated with the AM1 or PM3 semi-empirical method for all 106 chemicals. From 3D coordinates an extensive set of molecular descriptors was calculated. The classification model based on counterpropagation neural network (CP NN) was prepared and evaluated. The method of determining the thresholds necessary to convert the predictions from the CP NN into class-determinations, is described in details.

Keywords: *Artifical neural networks, QSPR study, endocrine disrupting chemicals*

Introduction

A large amount of chemicals is released every day in the environment having a large impact on human health and wildlife. Looking at environmental and occupational health issues, a four-stage process is investigated, covering source, exposure, tissue dose, and response, each of which requires the input and expertise of scientists and other professionals from many different disciplines. It is challenging to attack this interdisciplinary problem from different directions, having many disciplines routinely interact. Advances with cell and tissue cultures, computer modeling, and genetic research help to reduce the need for animals to test substances that can harm humanity, but the advances probably will not totally eliminate that need. In testing, computers allow toxicologists to develop mathematical models and algorithms that can predict the biological effects of new substances based on their chemical structure. If a new chemical has a structure similar to a known poison in certain key aspects, then the new substance also may be a poison. Such screening can thus preempt some animal use. Alternative methods are defined as methods, which replace the use of laboratory animals altogether, reduce the number of animals required, or refine existing procedures or techniques so as to minimize the level of stress endured by the animal.[1]

Toxicity is a broad definition of biochemical property. There are different mechanisms of action and different consequences of toxic effects of a given chemical.[2] Recent studies have shown that many toxic effects are based on the malfunctioning and disruption of the endocrine system.[3–6] Endocrine disrupters (EDs) are chemicals having capabilities to interfere with the endocrine systems. It is known for certain chemicals that they bind to the estrogen or androgen receptors. Most *in vitro* and *in vivo* data available on EDs in the literature are derived from assays that measure estrogenic or, less frequently, androgenic activity.

Computational or *in silico* methods,[7–12] alternative to *in vivo* and *in vitro* tests, are becoming essential because of the large amount of new chemicals emerging every day, and because of restrictions in ethically questionable animal tests. The European Union Commission reported about the candidate list of 553 substances that are potential endocrine disrupters.[13] The literature survey, about the substances suspected to act as endocrine disrupters, is given. Numerous information about data available in literature on several effects related to the endocrine disruption potency are extracted and grouped. A sub-set 106 compounds was chosen for further investigation, for the rest of compounds it was not possible to calculate structural descriptors needed for handling chemical structures, or the data on endocrine disruption activity was not reliable. The substances had been categorized into 3 stages of literature evidence for their endocrine disruption potency. Chemicals in the first category were confirmed to be endocrine

disrupters in an intact organism by at least one study found in the literature. The second category characterizes substances that are potentially active according to *in-vitro* studies, while the *in-vivo* data do not sufficiently prove the ED activity. For the third category, there was either no data available or data found for non scientific basis for inclusion into the list. Additional 244 substances were studied; however, the data from literature about their ED activity was less extensive or convincing.

There is a lack of homogeneity in data collected in the report,[13] because the individual studies refer to experiments made during decades, evaluating several species and a great variety of effects. It is difficult to select a specific feature, a numerical output, which would be a subject to be manipulated with the chemometrics techniques. Instead of modelling of certain biological endpoint, we decided to focus our research into the determination of a class of endocrine disruption activity (according to literature evidence) for individual compounds. The parameters that have to be input to the model are molecular structure descriptors, while the numerical output of the model contains the information to which class of endocrine disruption activity the compound belongs. The transformation of the numerical output of the model to the class-determination is based on a threshold value which has to be determined for each individual model and for each class separately.

Data

The database of 553 man-made chemicals suspected to act as endocrine disrupters was published by the EU Commission.[13] The chemicals were searched through the literature to find several effects related to the endocrine disruption potency. According to the report,[13] they were grouped in three categories: (1) Endocrine disrupter, (2) Potential endocrine disrupter, and (3) Non-active as endocrine disrupter. The third category was further split into Uncertainly-non-active and Non-active-endocrine-disrupter. See Table 1 for details.

After pruning the dataset (no literature data or molecular structure determinable), 106 structures (see Table 2) were accepted for the procedure of developing the automated classification model. All the structures were optimized with the MOPAC program, using AM1 or PM3 semi-empirical method to obtain 3D atomic co-ordinates. For a small group of compounds containing tin atoms (Sn) it was not possible to process this calculation with AM1 method. These compounds are printed in bold stamp in Table 2. The PM3 semi-empirical method, which provides parameterization also for Sn atoms, was applied in the case of tin-compounds.

Methods

The methods used to calculate descriptors of molecular structure were the following:

– MOPAC for 3D structure optimization (AM1 and PM3 semi-empirical methods for minimization of total molecular energy), to obtain atom co-ordinates.

Table 1 – *Categories of substances, suspected endocrine disrupting chemicals, studied by the EU Commission*

Category	Labeling	Description
1	**E**ndocrine disrupter	At least one study was found providing the evidence of endocrine disruption in an intact organism. Not a formal weight of evidence approach.
2	**P**otential endocrine disrupter	In vitro data indicating potential for endocrine disruption in intact organisms. Also includes effects *in-vivo* that may, or may not, be ED-mediated. May include structural analyses and metabolic considerations.
3	**U**ndefined activity or **N**on-endocrine disrupter	No scientific basis for inclusion in list of endocrine disrupters.
3A	**U**ndefined activity – No evidence for non-ED	No data available on wildlife relevant and/or mammal relevant endocrine effects.
3B	**U**ndefined activity – Some evidence for non-ED	Some data are available but the evidence is insufficient for identification.
3C	**N**on-endocrine disrupter – Certain evidence for non-ED	Data available indicating no scientific basis for inclusion into the list of active ED chemicals.

– CODESSA[14] for calculation of five classes of structural descriptors: constitutional, geometrical, topological, electrostatic, and quantum-chemical descriptors.

– Methods to obtain Log*P* values: experimental (from the experimental values database,[15] and from Hansch's manual[16]) or estimated by KowWin program.[17]

Counterpropagation neural network[18–20] was employed as a classification model. Below it is described shortly how the standard counterpropagation neural network can be modified to predict discrete classes of compounds. The counterpropagation neural network is based on a supervised learning method, only one part of the learning process (initial mapping of inputs) involves elements of the unsupervised learning. For the learning procedure a set of input-output pairs $\{\boldsymbol{X}_s, \boldsymbol{Y}_s\}$ is required. In the classification problem the input $\boldsymbol{X}_s = (x_{s1}, x_{s2}, \ldots x_{si} \ldots x_{sm})$ is a structure representation of the *s*-th compound, represented by *m* structural descriptors or "independent variables". The corresponding output or "dependent variables" $\boldsymbol{Y}_s = (t_{s1}, t_{s2}, \ldots t_{sj} \ldots t_{sp})$ is a *p*-component vector of zeros and ones. The value y_{sj} indicates whether the *s*-th compound is ($y_{sj} = 1$) or isn't ($y_{sj} = 0$) in the *j*-th class. The ANN is trained to respond for each input structure representation $\boldsymbol{X}_s$ from the training set with the output vector $\boldsymbol{Out}_s$ identical to the target (class-vector) $\boldsymbol{Y}_s$. The unsupervised element in the counterpropagation neural network learning procedure is the mapping of the structure-representation vectors into the

Table 2 – *The database of 106 compounds optimized with MOPAC, using AM1 or PM3 (compounds denoted by an asterisk) semi-empirical methods. The enumeration is taken from the complete set of compounds reported by the EU Commission.*[13]

No.	Class	Label	CASNR	Name
2	**2**	**P**	10605-21-7	Carbendazim
10	**2**	**P**	309-00-2	Aldrin
11	**1**	**E**	12789-03-6 (57-74-9)	Chlordane
13	**3 B**	**U**	3734-48-3	Chlordene
15	**2**	**P**	60-57-1	Dieldrin
16	**2**	**P**	115-29-7 (959-98-8 or 33213-65-9)	Endosulfan (also alfa and beta)
19	**2**	**P**	72-20-8	Endrin
20	**1**	**E**	143-50-0	Kepone (Chlordecone)
21	**1**	**E**	2385-85-5	Mirex
22	**2**	**P**	27304-13-8	Oxychlordane
25	**3 B**	**U**	39765-80-5	Trans-Nonachlor
27	**2**	**P**	94-75-7	2,4-Dichlorophenoxy acetic acid (2,4-D)
29	**2**	**P**	67747-09-5	Prochloraz
42	**1**	**E**	50-29-3	DDT (technical) = clofenotane = p,p'-DDT
44	**2**	**P**	115-32-2	Dicofol = Kelthane
57	**1**	**E**	3563-45-9	Tetrachloro DDT = 1,1,1,2--Tetrachloro-2,2-bis(4-chlorophenyl)ethane
60	**2**	**P**	36734-19-7	Iprodione
63	**1**	**E**	50471-44-8	Vinclozolin
73	**1**	**E**	137-26-8	Thiram
78	**1**	**E**	58-89-9	Gamma-HCH (Lindane)
85	**2**	**P**	330-54-1	Diuron
87	**1**	**E**	330-55-2	Linuron (Lorox)
104	**2**	**P**	333-41-5	Diazinon
106	**2**	**P**	60-51-5	Dimethoate
109	**3 C**	**N**	55-38-9	Fenthion
113	**2**	**P**	121-75-5	Malathion
115	**2**	**P**	298-00-0	Methylparathion
119	**2**	**P**	56-38-2	Parathion = Parathion(-ethyl)
141	**1**	**E**	61-82-5	Amitrol = Aminotriazol
142	**1**	**E**	1912-24-9	Atrazine
156	**2**	**P**	122-34-9	Simazine
159	**2**	**P**	43121-43-3	Triadimefon
163	**1**	**E**	34256-82-1	Acetochlor
164	**1**	**E**	15972-60-8	Alachlor
169	**3 A**	**U**	106-93-4	Dibromoethane (EDB)
176	**2**	**P**	76-44-8	Heptachlor
177	**3 B**	**U**	1024-57-3	Heptachlor-epoxide
179	**2**	**P**	74-83-9	Methylbromide (bromomethane)
182	**1**	**E**	1836-75-5	Nitrofen
183	**3 B**	**U**	4685-14-7	Paraquat = 1,1'-dimethyl-4,4'-bipyridinium
187	**2**	**P**	709-98-8	Propanil
190	**3 A**	**U**	29082-74-4	Octachlorostyrene
191	**1**	**E**	100-42-5	Styrene
194	**2**	**P**	120-83-2	2,4-Dichlorophenol
195	**2**	**P**	1570-64-5	4-chloro-2-methylphenol
196	**2**	**P**	59-50-7	4-chloro-3-methylphenol
198	**1**	**E**	118-74-1	Hexachlorobenzene (HCB)
215	**2**	**P**	98-54-4	4-*tert*-Butylphenol
216	**1**	**E**	140-66-9	4-*tert*-Octylphenol = 1,1,3,3-Tetramethyl-4-butylphenol
277	**3 B**	**U**	103-23-1	Bis(2-ethylhexyl)adipate
278	**1**	**E**	85-68-7	Butylbenzylphthalate (BBP)
279	**1**	**E**	117-81-7	Di-(2-ethylhexyl)phthalate (DEHP)
280	**3 B**	**U**	84-61-7	Dicyclohexyl phthalate (DCHP)
281	**3 B**	**U**	84-66-2	Diethyl phthalate (DEP)
283	**2**	**P**	26761-40-0	Diisodecyl phthalate
284	**2**	**P**	28553-12-0	diisononyl phthalate = 1,2-Benzenedicarboxylic acid, diisononyl ester (DINP)
286	**1**	**E**	84-74-2	Di-n-butylphthalate (DBP)

No.	Class	Label	CASNR	Name
318	**2**	**P**	1675-54-3	2,2'-bis(4-(2,3-epoxypropoxy)phenyl)propane = 2,2'-[(1- methylethylidene)bis(4,1-phenyleneoxymethylene)]bisoxirane
326	**1**	**E**	80-05-7	2,2-Bis(4-hydroxyphenyl)propan = 4,4'-isopropylidenediphenol = Bisphenol A
348	**3 A**	**U**	106-89-8	Epichlorohydrin (1-chloro-2,3-epoxypropane)
370	**3 B**	**U**	92-52-4	Diphenyl
371	**2**	**P**	90-43-7	o-phenylphenol
405	**3 B**	**U**	38380-07-3	PCB 128 (2,2',3,3',4,4'-Hexachlorobiphenyl)
406	**2**	**P**	38411-22-2	PCB 136 (2,2',3,3',6,6'-Hexachlorobiphenyl)
408	**1**	**E**	35065-27-1	PCB 153 (2,2',4,4',5,5'-Hexachlorobiphenyl)
409	**2**	**P**	38380-08-4	PCB 156 (2,3,3',4,4',5-Hexachlorobiphenyl)
410	**1**	**E**	32774-16-6	PCB 169 (3,3',4,4',5,5'-Hexachlorobiphenyl)
417	**1**	**E**	2437-79-8	PCB 47 (2,2',4,4'-Tetrachlorobiphenyl)
418	**2**	**P**	70362-47-9	PCB 48 (2,2',4,5-Tetrachlorobiphenyl)
419	**3 A**	**U**	35693-99-3	PCB 52 (2,2';5,5'-Tetrachlorobiphenyl)
420	**2**	**P**	33284-53-6	PCB 61 (2,3,4,5-Tetrachlorobiphenyl)
421	**2**	**P**	32598-12-2	PCB 75 (2,4,4',6-Tetrachlorobiphenyl)
422	**1**	**E**	32598-13-3	PCB 77 (3,3',4,4'-Tetrachlorobiphenyl)
435	**2**	**P**	No CAS 046	2,2',4,4'-Tetrabrominated diphenyl ether (2,2',4,4'-tetraBDE)
436	**2**	**P**	No CAS 044	Decabrominated diphenyl ether (decaBDE)
444	**3 B**	**U**	135-19-3	2-Naphthol
467	**1**	**E**	40321-76-4	1,2,3,7,8-Pentachlorodibenzodioxin
472	**1**	**E**	1746-01-6	2,3,7,8-Tetrachlorodibenzo-*p*-dioxin (2,3,7,8-TCDD)
483	**2**	**P**	57117-41-6	1,2,3,7,8-Pentachlorodibenzofuran
484	**2**	**P**	83704-53-4	1,2,3,7,9-Pentachlorodibenzofuran
485	**2**	**P**	58802-20-3	1,2,7,8-Tetrachlorodibenzofuran
486	**2**	**P**	71998-72-6	1,3,6,8-Tetrachlorodibenzofuran
487	**1**	**E**	57117-31-4	2,3,4,7,8-Pentachlorodibenzofuran (2,3,4,7,8-PeCDF)
488	**2**	**P**	67733-57-7	2,3,7,8-Tetrabromodibenzofuran
489	**2**	**P**	51207-31-9	2,3,7,8-Tetrachlorodibenzofuran
512*	**1**	**E**	688-73-3	Tributyltin hydride
513*	**1**	**E**	56-35-9	Tributyltin oxide = bis(tributyltin) oxide
516*	**1**	**E**	4342-30-7	Phenol, 2-[(tributylstannyl)oxy]carbony
517*	**1**	**E**	4342-36-3	Stannane, (benzoyloxy)tributyl-
518*	**1**	**E**	4782-29-0	Stannane, [1,2-phenylenebis(carbonyloxy)]
521*	**1**	**E**	24124-25-2	Stannane, tributyl[(1-oxo--9,12-octadecadienyl)]
522*	**1**	**E**	3090-35-5	Stannane, tributyl[(1-oxo-9-octadecenyl)]
524*	**1**	**E**	1983-10-4	Stannane, tributylfluoro-
525*	**1**	**E**	2155-70-6	Tributyl[(2-methyl-1-oxo--2-propenyl)oxy]stannane
530*	**1**	**E**	1461-25-2	Tetrabutyltin (TTBT)
531*	**1**	**E**	668-34-8	Triphenyltin
532*	**1**	**E**	900-95-8	Fentin acetate = triphenyltin acetate
536	**1**	**E**	95-76-1	3,4-Dichloroaniline
538	**1**	**E**	99-99-0	4-Nitrotoluene
541	**3 A**	**U**	119-61-9	Benzophenone
545	**3 A**	**U**	68-12-2	Dimethylformamide (DMFA)
548	**3 C**	**N**	107-21-1	Ethylene glycol (ethane-1,2-diol)
557	**2**	**P**	127-18-4	Perchloroethylene
558	**3 C**	**N**	108-95-2	Phenol
560	**1**	**E**	108-46-3	Resorcinol
564	**3 B**	**U**	108-05-4	Vinyl acetate

Kohonen layer (input layer of the counterpropagation neural network consisting of $n_x \times n_y$ neurons). For this step no knowledge about the target vector is needed. Once the position of the input vector is defined, the weights of the neurons in, both, input and output layers are corrected according to the particular element from the training set, $\{\boldsymbol{X}_s, \boldsymbol{Y}_s\}$ pair (training object). The trained output layer consists of $n_x \times n_y$ output neurons arranged in squared neighborhood. The levels of the output layer represent p response surfaces for the p classes. The points of the response surfaces correspond to the weights of the output neurons $\boldsymbol{Out} = (out_1, out_2, \ldots out_j \ldots out_p)$. After the training, each weight out_j is a numerical value between 0.0 and 1.0. For the final prediction of classes the response surface values must be again transformed into discrete values, zeros and ones. The threshold value, between 0.01 and 0.99, must be determined for each of the p classes. Below the threshold all predictions are negative and denoted by a zero, what means that the s-th compound does not belong to the j-th class, while the predictions above the threshold are positive and denoted by one. The threshold is determined according to the number of correct/wrong class predictions if the trained network is tested by the same objects as it was trained with, i.e. $\{\boldsymbol{X}_s, \boldsymbol{Y}_s\}$ pairs from the training set.

Results and discussion

The classification model was developed and tested on the dataset containing 106 compounds ($N_{mol} = 106$). The molecular structures were described by constitutional, topological, geometrical, electrostatic and quantum-chemical descriptors calculated with CODESSA. 766 descriptors were obtained; 484 of them were available only for a limited number of molecules (so called incomplete descriptors), while 16 descriptors were equal for all molecules and thus neglected. The remaining 266 descriptors of each molecular structure ($m = 266$) were descriptive and available for all compounds, thus accepted for structural descriptors. The descriptors calculated by CODESSA from molecular 3D co-ordinates were appended by an experimentally obtained parameter Log*P*, which reflects the compounds' hydrophobic property usually playing an important role in the mechanism of action of particular biological activity.[21] Log*P*, the logarithm of octanol-water partition coefficient, describes equilibrium partitioning of a chemical between octanol and water phases. Experimental Log*P* values were obtained from literature (evidence from Physical Properties Database[15]), from *Hansch*,[16] or estimated using KowWin program.[17] All descriptors were normalized with mean = 0 and standard deviation = 1.

The ED categories associated with the 106 compounds from the dataset are following:

– Category No. 1 with label E (evidently active) – 43 compounds

– Category No. 2 with label P (potentially active) – 43 compounds

– Category No. 3 (A+B) with label U (uncertain evidence) – 17 compounds

– Category No. 4 (C) with label N (non active) – 3 compounds

The literature evidence of the risk for a chemical to be an endocrine disrupter is decreasing from the first towards the fourth class. The first 3 classes ($p = 1\ldots3$) are relatively well populated, while there is evident lack of compounds in the fourth ($p = 4$) class. We decided to split the third category into two classes, because the uncertain evidence of 3A and 3B is not strong enough for such an important decision, which our predictive model is trained for, that would classify a chemical to be harmless regarding the endocrine disrupting activity. Only for the category 3C there is no drought about non-activity.

For the model building purpose the data was split into the training and the test set using the Kohonen maps as the selection method.[20,22] Since the compounds are not evenly distributed between the classes, we made the selection in such a way, that two thirds of compounds of each class were kept for training, while one third for testing and validating the constructed classification model. 71 compounds were assigned to the training set, 35 to the test set.

A method similar to the one used for the selection of training and test sets[20,22] was applied for the selection of descriptors. The main difference is in the way how the matrix of input data is represented; in the case of the selection of descriptors the transposed data matrix is used instead of original data matrix, in which the rows and columns ($N_{mol} = 106$ rows and $m = 266$ columns) correspond to molecules and descriptors, respectively. The transposed matrix consists of $m = 266$ rows (descriptors) and $N_{mol} = 106$ columns (molecules). It is important that the transposed matrix is normalized column-wise before it is used for training the Kohonen network for a certain training time (epochs). The result is a Kohonen map, in which the descriptors are self-organized onto the $n_x \times n_y$ positions (neurons). A Kohonen network with $5 \times 5 = 25$ neurons was used producing a map with 25 positions. All 266 descriptors were placed onto these 25 positions (neurons). This means that each neuron was occupied in average by 11 descriptors. In Fig. 1 it is demonstrated how many descriptors were placed on individual neurons.

The neurons and descriptors are labeled by the indices i and s, respectively. In the procedure for selecting descrip-

Ny \ Nx	1	2	3	4	5
1	*12*	*10*	*4*	*9*	*11*
2	*11*	*9*	*3*	*10*	*6*
3	*4*	*8*	*14*	*3*	*6*
4	*22*	*20*	*3*	*8*	*10*
5	*10*	*7*	*4*	*25*	*37*

Fig. 1 – *The distributions of descriptors in the 5 × 5 top-map of the Kohonen neural network. (a) Number of descriptors occupying an individual neuron; (b) Top left section of the top-map with a list of descriptors on the neurons shown; (c) Two descriptors from each neuron chosen on the basis of smallest and largest distance between the neuron and the descriptor's vector.*

tors, the dimension of neurons is equal to the number of molecules in the data-set (N_{mol} = 106); this is also the number of components of the descriptors' representation vectors obtained as rows in the transposed data matrix ($\boldsymbol{X}^T_{js}$, $j = 1, N_{mol}$). The i-th neuron is represented as a vector of weights ($\boldsymbol{W}_{ji}$, $j = 1, N_{mol}$). In the training procedure of the Kohonen neural network, similar descriptors are falling onto the neighboring neurons. If the number of training objects significantly exceeds the number of neurons, many objects occupy the same neuron. The criterion for the selection of descriptors assembled on the same neuron was the Euclidean distance between the descriptor $\boldsymbol{X}_s^T$ and neuron $\boldsymbol{W}_i$:

$$d_{s,i} = \sqrt{\sum_{j=1}^{N_{mol}} (\boldsymbol{X}^T_{j,s} - \boldsymbol{W}_{j,i})^2} \qquad (1)$$

Only two descriptors from each neuron were chosen for final representation of molecular structure, one with the smallest and one with the largest distance from the excited neuron.

The network with dimensions 5 × 5 × 106 was trained for 50, 100, 300, 500, and 1000 epochs. The distribution of objects (descriptors) in the 5 × 5 top-map and the distances of all objects on one neuron was examined. The network trained for 300 epochs was chosen for final descriptor selection procedure because of the most even distribution of objects and small differences between the maximal and minimal distances $d_{s,i}$ calculated at each neuron. The reduced set contained 50 descriptors, two from each neuron: the most similar one and most different one regarding the distance from the particular neuron (Eq. 1).

Two different sets of descriptors were tested for this study: the non-reduced set of 266 descriptors, and the reduced set of 50 descriptors, for which the reduction method is described above. With these two datasets, each divided into a training (72 molecules) and test set (35 molecules), different models were built. The CP NN parameters that were varied were: number of neurons ($n_x \times n_y$), training time (epochs), while the learning rate and momentum term were 0.5 and 0.01, respectively, in all constructed models.

The evaluation of the class-predictions from the resulting models is not straightforward. The predictions are obtained from the output layers of individual models. However, they are presented as real numbers from 0.0 to 1.0, one prediction from each of the four levels of the output layer for four possible classes. As described in the **Methods** section, the prediction obtained for individual molecule is a four-dimensional vector $\boldsymbol{Out} = (out_1, out_2, out_3, out_4)$,$D$ $0.00 \le out_j \le 1.00$. It is necessary to determine the threshold value (T^+), above which the prediction for a j-th class is positive (confirmative). T^+ enables the transformation of the model output values to discrete class predictions, i.e. one for a confirmative and zero for a rejecting answer. There are four classes, so we need four threshold values for each of the constructed model (T_j^+ , j = 1, 4). They are determined according to the number of correct/wrong class predictions if the trained network is tested by the same objects as it was trained with, i.e. molecules from the training set. Below the threshold all predictions are rejecting and denoted by a zero, what means that the compound does not belong to the j-th class, while the predictions above the threshold are positive and denoted by one (the compound belongs to the j-th class). In Fig. 2 to 4 examples of the determination of T_j^+ for three different constructed models are shown.

As can be seen from Figs. 2–4, the individual threshold is positioned where the sum of errors of, both, false positive and false negative predictions, is the lowest. If the T_j^+ were positioned close to zero, the predictions of the j-th class for most of the molecules from the training set would be con-

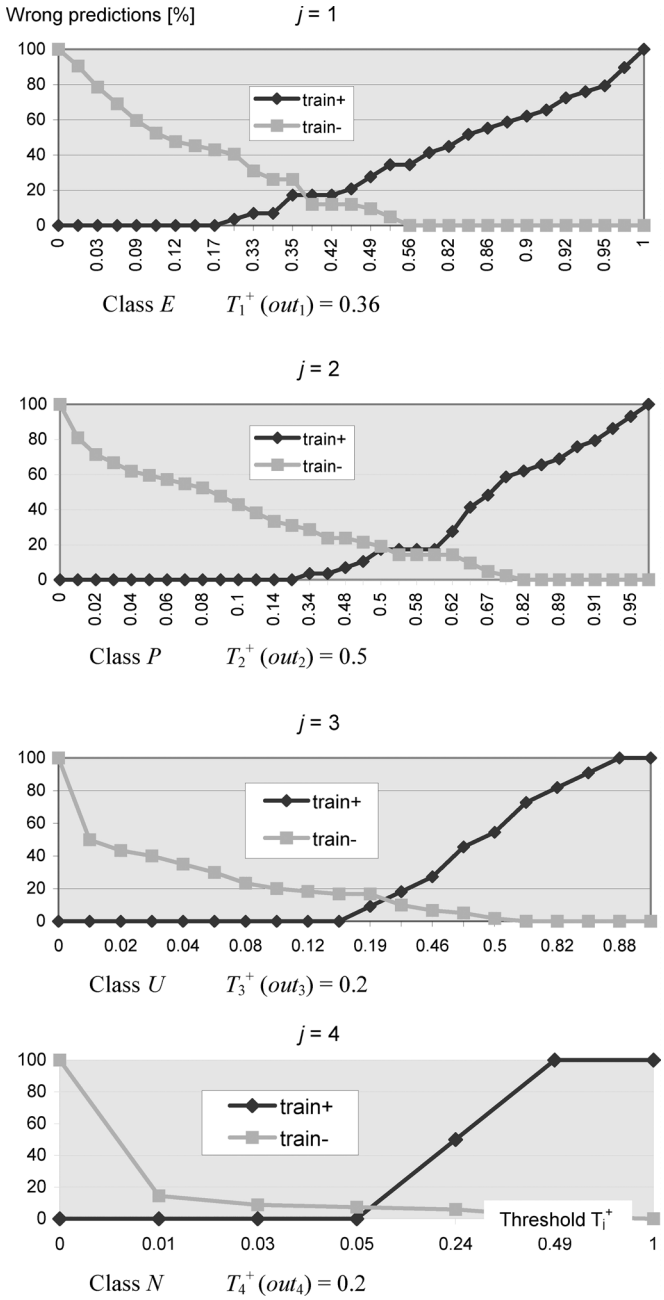

Fig. 2 – *The thresholds determined T_j^+ for the class-predictions in the model from the counterpropagation neural network of 9 x 9 neurons, trained for 100 epochs with the molecules represented by a non-reduced set of descriptors. The diamonds and squares stand for positive (confirmative) and negative (rejecting) predictions, respectively.*

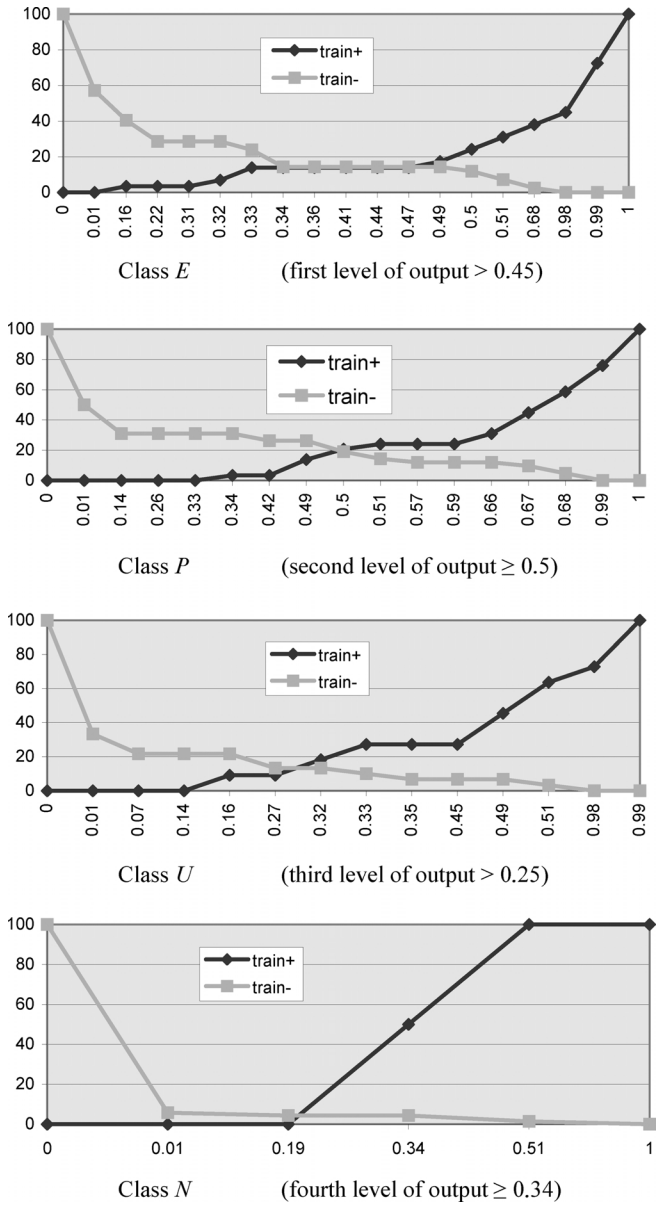

Fig. 3 – *The thresholds determined T_j^+ for the class-predictions in the model from the counterpropagation neural network of 9 x 9 neurons, trained for 300 epochs with the molecules represented by a non-reduced set of descriptors. The diamonds and squares stand for positive (confirmative) and negative (rejecting) predictions, respectively.*

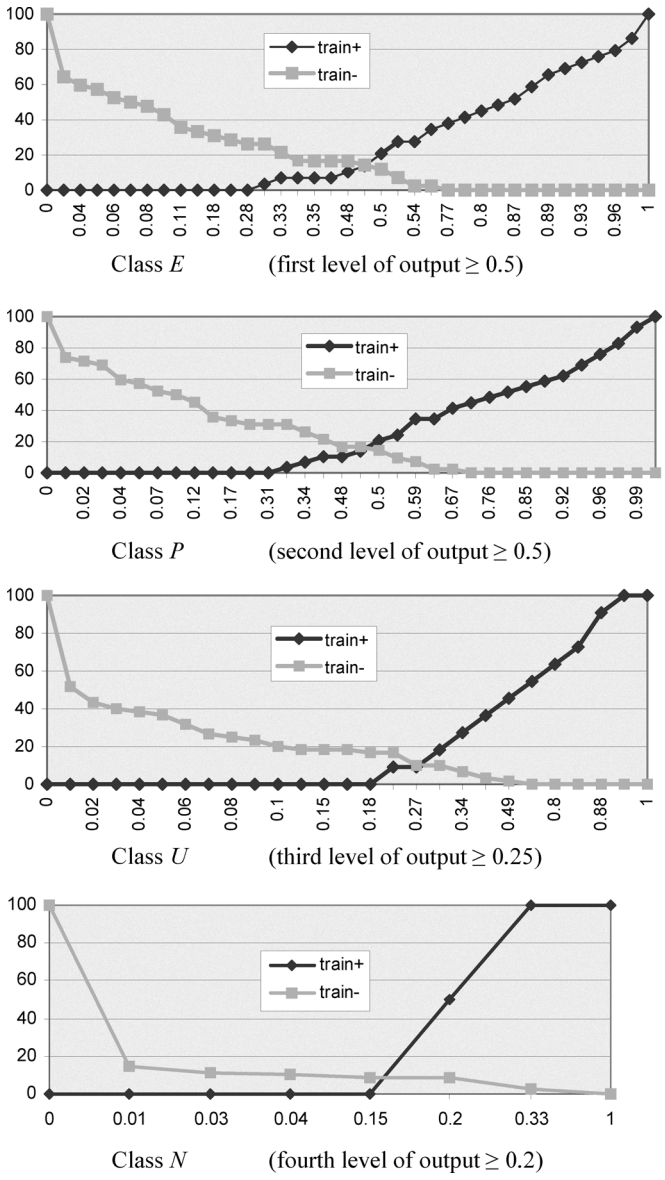

Fig. 4 – *The thresholds determined T_j^+ for the class-predictions in the model from the counterpropagation neural network of 9 x 9 neurons, trained for 100 epochs with the molecules represented by a reduced set of 50 descriptors. The diamonds and squares stand for positive (confirmative) and negative (rejecting) predictions, respectively.*

firmative ($\boldsymbol{Out}_j > 0$). The molecules from the *j*-th class would be correctly predicted, while the predictions for the rest of molecules would be so called false positive. On the other hand, if the T_j^+ were close to one, majority of predictions for class *j* would be rejecting. This would produce false negative predictions of the molecules that are actually in the *j*-th class. The threshold has to be determined for each individual model when tested for its predictive ability.

Once the thresholds were defined, the models were validated by checking the class-predictions for 35 test molecules. The misclassification tables, obtained by comparison of actual and predicted classes of test compounds, are shown in Fig. 5.

The best model is chosen on the basis of several criteria:

– the largest number of correct predictions (sum of the diagonal elements);

– the smallest number of false negative predictions, which are more severe errors than false positives, because they would classify a harmful compound as a nontoxic one;

– the smallest sum of predictions that are wrong for more then one category (model (k) in Fig. 5).

Model (a) from Fig. 5, with 69 % of correct predictions, would be the best if used for class-predictions, while for the priority settings model (k) is better, because it makes the range-list of tested chemicals from most to least harmful (according to the literature evidence) less erroneous, because the prediction never misses the correct class for more than one class.

(a) DS1; 9 x 9; 100 epochs

PRED. \ TRUE	E	P	U	N
E	11	1	2	1
P	2	12	3	0
U	1	1	1	0
N	0	0	0	0

(e) DS2; 9 x 9; 100 epochs

PRED. \ TRUE	E	P	U	N
E	10	4	2	0
P	3	9	3	0
U	1	1	1	1
N	0	0	0	0

(i) DS3; 9 x 9; 100 epochs

PRED. \ TRUE	E	P	U	N
E	12	5	3	1
P	2	9	3	0
U	0	0	0	0
N	0	0	0	0

(b) DS1; 9 x 9; 300 epochs

PRED. \ TRUE	E	P	U	N
E	10	4	1	1
P	4	9	3	0
U	0	1	2	0
N	0	0	0	0

(f) DS2; 9 x 9; 300 epochs

PRED. \ TRUE	E	P	U	N
E	9	2	0	1
P	4	11	3	0
U	1	1	3	0
N	0	0	0	0

(j) DS3; 9 x 9; 300 epochs

PRED. \ TRUE	E	P	U	N
E	10	5	1	1
P	3	7	4	0
U	1	2	1	0
N	0	0	0	0

(c) DS1; 12 x 12; 100 epochs

PRED. \ TRUE	E	P	U	N
E	13	4	2	1
P	1	9	3	0
U	0	1	1	0
N	0	0	0	0

(g) DS2; 12 x 12; 100 epochs

PRED. \ TRUE	E	P	U	N
E	10	3	2	1
P	3	11	2	0
U	1	0	2	0
N	0	0	0	0

(k) DS3; 12 x 12; 100 epochs

PRED. \ TRUE	E	P	U	N
E	10	4	0	0
P	4	10	5	0
U	0	0	1	1
N	0	0	0	0

(d) DS1; 12 x 12; 300 epochs

PRED. \ TRUE	E	P	U	N
E	10	1	2	1
P	3	12	3	0
U	1	1	1	0
N	0	0	0	0

(h) DS2; 12 x 12; 300 epochs

PRED. \ TRUE	E	P	U	N
E	9	4	1	1
P	4	8	4	0
U	1	1	1	0
N	0	0	0	0

(l) DS3; 12 x 12; 300 epochs

PRED. \ TRUE	E	P	U	N
E	10	2	1	0
P	3	12	5	0
U	1	0	0	1
N	0	0	0	0

Fig. 5 – *Classification tables with the number of correct (diagonal elements), false positive (upper triangle), and false negative predictions (lower triangle). The predictions are acquired from 12 models (from (a) to (l)), constructed on the basis of three different spectral representations (DS1, DS2 and DS3), using two different neural network architectures (9 x 9 and 12 x 12 neurons), while the training time was 100 or 300 epochs.*

Conclusions

A computational model based on the counterpropagation neural networks for classification of endocrine disrupter activity of compounds of known chemical structures, is proposed. The emphasis is on the determination of the threshold for each model, which converts the real number predictions into a discrete class number. The dataset contains structurally very diverse chemicals. Nevertheless, the two-step modelling principle of the counterpropagation neural network enables to build a classification model capable of treating all chemicals together. The class predictive power of constructed models is reasonable for priority setting and would be significantly improved if more data were available, specially in the low endocrine activity region.

ACKNOWLEDGEMENT

Authors acknowledge partial financial support of the Ministry of Education, Science, and Sport of Slovenia for financing the research through the project grants P104-507, as well as the European Union 5-th framework program scheme for partial financial support through the program Marie Curie Host Training Site no. HPMT-CT-2001-00240 and IMAGETOX no: HPRN-CT-1999-00015.

References

1. *A. N. Rowan,* Of Mice, Models, & Men: A Critical Evaluation of Animal Research Albany, NY: State University of New York Press, 1984.
2. *U. A. Boelsterli,* Mechanistic Toxicology, The Molecular Basis of How Chemicals Disrupt Biological Targets, Taylor & Francis, New York, 2002.
3. *T. Colborn, F. S. V. Saal, A. M. Soto,* Environmental Health Perspectives **101** (1993) 378.
4. *T. H. Hutchinson, D. B. Pickford,* Toxicology **181** (2002) 383.
5. *B. Brunstrom, J. Axelsson, K. Halldin,* Ecotoxicology **12** (2003) 287.
6. *D. W. Singleton, S. A. Khan,* Frontiers in Bioscience **8** (2003) 110.
7. *T. W. Schultz, M. T. D. Cronin, T. I. Netzeva,* Journal of Molecular Structure-Theochem **622** (2003) 23.
8. *S. P. Bradbury, C. L. Russom, G. T. Ankley, T. W. Schultz, J. D. Walker,* Environmental Toxicology and Chemistry **22** (2003) 1789.
9. *A. D. P. Worgan, J. C. Dearden, R. Edwards, T. I. Netzeva, M. T. D. Cronin,* QSAR & Combinatorial Science **22** (2003) 204.
10. *M. T. D. Cronin, J. D. Walker, J. S. Jaworska, M. H. I. Comber, C. D. Watts, A. P. Worth,* Environmental Health Perspectives **111** (2003) 1376.
11. *M. T. D. Cronin, T. W. Schultz,* Journal of Molecular Structure-Theochem **622** (2003) 39.
12. *D. R. J. Moore, R. L. Breton, D. B. MacDonald,* Environmental Toxicology and Chemistry **22** (2003) 1799.
13. European Commission, Communication from the Commission to the Council and the European Parliament on the Implementation of the Community Strategy for Endocrine Disrupters – a Range of Substances Suspected of Interfering with the Hormone Systems of Humans and Wildlife, COM (2001) 262 final, Brussels, 14 June 2001.
14. *A. R. Katritzky, V. S. Lobanov, M. Karelson,* CODESSA 2.0, Comprehensive Descriptors for Structural and Statistical Analysis, Copyright© 1994–1996 University of Florida, U.S.A.
15. Physical Properties Database – PHYSPROP – http://esc.syrres.com/interkow/PhysProp.htm
16. *L. Hansch, A. Leo,* Exploring QSAR Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, DC 1995
17. KowWin program v1.66, on-line demo: http://esc.syrres.com/interkow/kowdemo.htm
18. *R. Hecht-Nielsen,* Appl. Optics **26** (1987) 4979.
19. *J. Dayhof,* Neural Network Architectures, An Introduction; Van Nostrand Reinhold, New York, 1990.
20. *J. Zupan, M. Novič, I. Ruisanchez,* Chemometr. Intell. Lab. Syst. **38** (1997) 1.
21. *C. Hansch, D. Hoekman, A. Leo, D. Weininger, C. D. Selassie*, Chemical Reviews **102** (2002) 783.
22. *M. Novič, J. Zupan,* J. Chem. Inf. Comput. Sci. **35** (1995) 454.

SAŽETAK

Primjena umjetnih neuralnih mreža u QSPR istraživanju – Automatska klasifikacija kemikalija štetnih za endokrini sustav

*M. Novič i A. Roncaglioni**

Europska unija je dostavila popis od 553 kemikalije koje se trebaju ispitati radi mogućih štetnih djelovanja. U izvješću temeljenom na desetgodišnjem eksperimentiranju procijenjen je niz učinaka koji pokazuju nehomogenost dobivenih podataka. Na temelju objavljenih činjenica o njihovom djelovanju, Komisija je predložila klasifikaciju oštećivača endokrinog sustava (EDs). U ovom prilogu prikazuje se prijedlog metodologije kojom bi se pronašao model za automatsko predviđanje pripadnosti pojedinim kategorijama. Za rješenje tog problema primijenjene su tehnike skupljanja i klasifikacije. Iz popisa od 553 kemikalije, za 106 molekula s određenom kemijskom strukturom određena je pripadnost ED klasi. Molekulske strukture svih 106 kemikalija prikazane su pomoću 3D atomskih koordinata izračunatih AM1 ili PM3 semiempirijskim metodama. Iz 3D koordinata izračunati su molekulski deskriptori. Ispitan je klasifikacijski model koji se temelji na neuralnim mrežama CP NN (counterpropagation neural network).

Kemijski nacionalni inštitut, Ljubljana, Slovenija
** Institut Mario Negri, Milano, Italija*

Prispjelo 12. prosinca 2003.
Prihvaćeno 18. svibnja 2004.