Counter propagation artificial neural network categorical models for prediction of carcinogenicity for non-congeneric chemicals

Article *in* SAR and QSAR in environmental research · January 2010

DOI: 10.1080/10629360903563250 · Source: PubMed

4 authors, including:

Natalja Fjodorova
National Institute of Chemistry
40 PUBLICATIONS **529** CITATIONS

SEE PROFILE

Marjan Vracko
National Institute of Chemistry
156 PUBLICATIONS **3,634** CITATIONS

SEE PROFILE

Marjana Novič
National Institute of Chemistry
245 PUBLICATIONS **4,582** CITATIONS

SEE PROFILE

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

Counter propagation artificial neural network categorical models for prediction of carcinogenicity for non-congeneric chemicals

N. Fjodorova [a]; M. Vračko [a]; A. Jezierska [ab];M. Novič [a]
[a] National Institute of Chemistry, Ljubljana, Slovenia [b] Faculty of Chemistry, University of Wrocław, Wrocław, Poland

Online publication date: 06 April 2010

To cite this Article Fjodorova, N. , Vračko, M. , Jezierska, A. andNovič, M.(2010) 'Counter propagation artificial neural network categorical models for prediction of carcinogenicity for non-congeneric chemicals', SAR and QSAR in Environmental Research, 21: 1, 57 — 75
To link to this Article: DOI: 10.1080/10629360903563250
URL: http://dx.doi.org/10.1080/10629360903563250

SAR and QSAR in Environmental Research
Vol. 21, Nos. 1–2, January–March 2010, 57–75

Taylor & Francis
Taylor & Francis Group

Counter propagation artificial neural network categorical models for prediction of carcinogenicity for non-congeneric chemicals†

N. Fjodorova[a]*, M. Vračko[a], A. Jezierska[a,b] and M. Novič[a]

[a]*National Institute of Chemistry, Ljubljana, Slovenia;* [b]*Faculty of Chemistry, University of Wrocław, Wrocław, Poland*

(*Received 6 July 2009; in final form 17 November 2009*)

One of the main goals of the new chemical regulation REACH (Registration, Evaluation and Authorization of Chemicals) is to fill the gaps on the toxicological properties of chemicals that affect human health. Carcinogenicity is one of the endpoints under consideration. The information obtained from (quantitative) structure–activity relationship ((Q)SAR) models is accepted as an alternative solution to avoid expensive and time-consuming animal tests. The reported results were obtained within the framework of the European project 'Computer Assisted Evaluation of industrial chemical Substances According to Regulations (CAESAR)'. In this article, we demonstrate intermediate results for counter propagation artificial neural network (CP ANN) models for the prediction category of the carcinogenic potency using two-dimensional (2D) descriptors from different software programs. A total of 805 non-congeneric chemicals were extracted from the Carcinogenic Potency Database (CPDBAS). The resulting models had prediction accuracies for internal (training) and external (test) sets as high as 91–93% and 68–70%, respectively. The sensitivity and specificity of the test set were 69–73 and 63–72% correspondingly. High specificity is critical in models for regulatory use that are aimed at ensuring public safety. Thus, the errors that give rise to false negatives are much more relevant. We discuss how we can increase the number of correctly predicted carcinogens using the correlation between the threshold and the values of the sensitivity and specificity.

Keywords: REACH; QSAR; CP ANN; categorical models; ROC; carcinogenicity

1. Introduction

The evaluation of chemical toxicity with respect to human health risk is of primary interest because it is connected to current regulatory actions regarding new and existing chemicals. It is well known that full implementation of the European chemical regulation REACH (Registration, Evaluation and Authorization of Chemicals) would require testing of around 30,000 existing substances. On the one hand, this is expensive and requires animal testing, whereas, on the other hand, the so-called 3Rs policy of replacing, reducing and refining the use of animal tests requests the development of alternatives to animal testing methods. Among the likely alternatives are quantitative structure–activity relationship (QSAR) methods [1].

*Corresponding author. Email: natalja.fjodorova@ki.si
†Presented at CMTPI 2009: Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (Istanbul, Turkey, 4–8 July, 2009).

ISSN 1062–936X print/ISSN 1029–046X online

DOI: 10.1080/10629360903563250
http://www.informaworld.com

Carcinogenicity is one of the most essential endpoints in the assessment of human health safety of chemicals. Many models for prediction of carcinogenic potency have been published in recent years [2–6]. Some QSAR models have been developed for particular chemical classes (such as amines, nitro compounds, polycyclic aromatic hydrocarbons) [7–9]. A large number of expert systems have been developed for prediction of carcinogenicity. Some of them are based on different endpoints and their combinations (so-called integrated systems) [10–17]. Models for non-congeneric chemicals are of great interest for regulatory use because they involve various classes of chemicals [18,19]. Frameworks, state of the art and perspectives of predictive models for carcinogenicity as well as mutagenicity have been described in recent paper [20]. It was pointed out that [20,21]:

> ... good local QSARs for congeneric chemicals can attain 70–100% correct external predictions if they are used to discriminate between inactive and active (mutagens, carcinogens) chemicals. This result indicates that these QSARs can be used with good reliability for applicative purposes (e.g., enriching the target for priority setting).

For non-congeneric chemicals, it was accepted an accuracy of the external test set or percentage of chemicals correctly predicted in external dataset equal to 65% [22].

The big challenge in carcinogenicity prediction is to construct a model that is able to predict carcinogenicity for a wide diversity of molecular structures, spanning an undetermined number of chemical classes and biological mechanisms. Among the statistical approaches for prediction of complex endpoints such as carcinogenicity, artificial neural networks (ANNs) appear to be one of the most suitable and promising for large datasets of chemicals. Compared to expert systems where chemical data are handled in several formats, the application of neural networks employs molecular descriptors, which indeed have been used in the prediction of carcinogenicity with contrasting results [23–25]. ANNs stand out from other machine-learning techniques because of their perceived ability to mimic activities of the human brain [26]. Gasteiger and Zupan [27] published the first fundamental description about the use of ANNs in chemistry. It should be highlighted that many interesting articles in this field have been published [28–30]. The contemporary applications of ANNs in life sciences have been extensively reviewed [31–33]. ANNs have found widespread use for classification tasks, function approximation and non-linear modelling, clustering and prediction in many fields of chemistry and bioinformatics [33–35]. The main advantage of neural network modelling is that the complex, non-linear relationships can be modelled without any assumptions about the form of the model. Large datasets can be examined. Neural networks are able to cope with noisy data and are fault-tolerant. Among the features of ANNs that could be considered as disadvantages are the fact that they function largely as a black box and understanding of the acquired knowledge is not always possible [36].

In the article, we describe QSAR models for non-congeneric chemicals for the prediction of carcinogenic potency using the counter propagation (CP) ANN method. Our models were developed in accordance with the principles of validation adopted by the Organization for Economic Co-operation and Development (OECD) in the scope of the European Commission (EC) funded project 'Computer Assisted Evaluation of industrial chemical Substances According to Regulation (CAESAR)' [37].

In silico methods are used in risk assessment for priority setting, mechanistic studies and others purposes [20].

The goal of the present article is to show intermediate results for the carcinogenicity models obtained in the CAESAR project. We have described only CP ANN models for the

prediction category of the carcinogenic potency using two-dimensional (2D) descriptors from the MDL, DRAGON and CODESSA software programs.

In the case of models for regulatory purposes, it is important to ensure public safety. Therefore, in this paper we examine how one can increase the number of correctly predicted carcinogens using a correlation between the threshold of the categorical models and the sensitivity and specificity. We address the issue of threshold effects on the overall performance of the models.

2. Materials and methods

2.1 *Data*

In this study we used the Carcinogenic Potency Database (CPDBAS) summary tables version 3b, updated 10 April 2006, obtained from the Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html [38]. These tables show summarized results for experiments on 1481 substances. CPDBAS is based on data taken from the Lois Gold Carcinogenic Potency Database (CPDB) http://potency.berkeley.edu/cpdb.html [39]. CPDBAS is an example of integration of data provided by collaboration among researchers involved in:

- the Distributed Structure-Searchable Toxicity (DSSTox) project;
- the Carcinogenic Potency Project;
- projects at the National Cancer Institute; and
- the PubChem database.

The carcinogenicity potency dataset, employed in our study, includes 805 compounds extracted from CPDBAS version 3b. The full list of 805 chemicals is available in Table S1 of the supplementary material which is available via the supplementary content tab on the online article webpage. All incorrect structures, ambiguous or mixed structures, polymers, inorganic compounds, metallo-organic compounds, salts, complexes and compounds without well-defined structure were eliminated from the initial datasets of 1481 chemicals. The carcinogenic potency for rats (males and females) was selected as the response, because such data in risk assessment [40] are often considered more suitable for human carcinogenicity prediction. The obtained data were cross-checked by at least two of the partners involved in the CAESAR project and were then used for descriptor calculation and mathematical modelling.

2.2 *Composition of the training and test sets*

The dataset of 805 chemicals was subdivided into training (644 chemicals) and test (161 chemicals) sets using the sub-sorting of chemicals according to functional groups and the following procedure was then aimed to distinguish between the connectivity aspects. At first, the chemicals were sorted according to a hierarchical system of compound classes with respect to functional groups. Next, within compound classes the compounds were sorted according to halogen or aromatic substitution, bond order, ring contents and finally according to the chemical formula (i.e., the number of atoms of different types). The sorting of chemicals was made in such a way that, in each subset (training or test), all major structural features were represented according to their relative occurrences in the total compound set.

This part of the study was carried out at the Helmholtz Centre for Environmental Research – UFZ in Germany by one of the groups involved in the CAESAR project. The sorting of the compounds was implemented in the software system ChemProp [41,42].

Analysis of the distribution of carcinogens and non-carcinogens in the total, training and test sets yielded the following results. The total dataset (805 chemicals) contains 422 carcinogens and 383 non-carcinogens. The training (644 chemicals) and test (161 chemicals) sets contain 327 and 95 carcinogens and 317 and 66 non-carcinogens, respectively. It is worth highlighting the fact that positive (carcinogens) and negative (non-carcinogens) compounds are evenly distributed over all of the examined sets.

It should be noticed that in this study that carcinogens were classified as active (P–positive) and non-carcinogens were classified as inactive (NP–not positive) compounds.

2.3 *Generation and selection of descriptors*

Currently various sets of molecular descriptors are available [43]. Different software packages for calculation of the descriptors have been developed and described [44–47]. In this study, we generated the following sets of descriptors: 254 MDL descriptors calculated by MDL QSAR version 2.2 [46], 835 DRAGON descriptors calculated by DRAGON Professional 5.4 [47] and 88 CODESSA descriptors calculated using CODESSA version 2.21 [45]. The next step in the study was the reduction in the number of descriptors and selection of the most informative of those for the carcinogenicity prediction.

Variable selection is an important issue in quantitative structure–activity/property relationship modelling. Nowadays, it is possible to generate hundreds of descriptors belonging to different classes such as the constitutional, topological, topochemical, topographical, geometrical or quantum-chemical classes [48,49], but the following question then arises: Which of them are the most significant for correlation with biological activity or other analysed properties? The literature study showed us that variable selection is a topic that has been intensively investigated over last few years. Many approaches have been developed and reported as tools for this purpose [50–58].

The goal of our study was to reduce the descriptor 'noise' termed as feature selection. Different partners of the consortiums involved in the CAESAR project employed different methods and techniques. The Central Science Laboratory (CSL), UK proposed techniques for selection of descriptors which was based on a cross-correlation matrix, multi-collinearity technique, fisher ratio and genetic algorithm. To determine the most important variables for model prediction the National Institute of Chemistry Ljubljana, Slovenia applied the Kohonen neural network (KNN) and principal component analysis (PCA) [58–62]. All descriptors were auto-scaled (e.g. normalized with zero mean and standard deviation equal to one).

2.4 *Counter propagation artificial neural network*

A CP ANN was employed in our study to develop the classification (categorical) models. The architecture of the CP ANN is presented in Figure 1.

In a general way, the CP ANN can be explained as follows. The input or Kohonen layer contains information on the input values which are vector representing structure (Figure 1). For example, the structure of the sth compound represented by m structural descriptors or 'variables' can be expressed as $\boldsymbol{X}_s = (x_{s1}, x_{s2}, \ldots, x_{si}, \ldots, x_{sm})$.

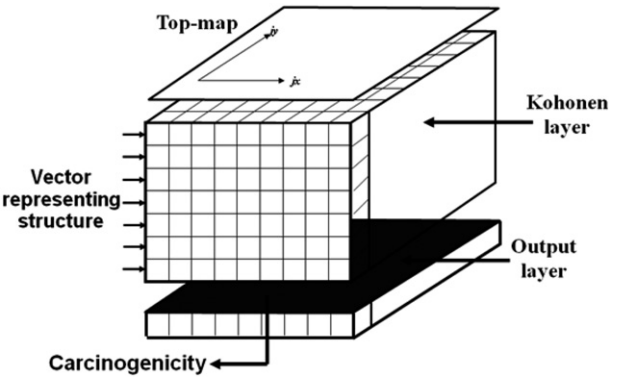

Figure 1. Counter propagation artificial neural network (CP ANN) architecture.

The output layer is associated with the output values, the so-called target $\boldsymbol{T}_s = (t_{s1}, t_{s2}, \ldots, t_{sj}, \ldots, t_{sp})$, which is a p-component vector of zeros and ones. The target in our classification model expresses the carcinogenicity class (P–positive = 1 and NP–not positive = 0). For each input structure representation $\boldsymbol{X}_s$ from the training set, the neural network is trained to respond with the output vector $\boldsymbol{Out}_s$ identical to the target (class-vector) $\boldsymbol{T}_s$. The Kohonen input layer of the CP ANN consists of $n_x \times n_y$ neurons. After the learning procedure, the objects are organized in such a way that similar objects are situated close to each other. We emphasize that only the input values participate in this phase of the learning (the unsupervised step). For this step, no knowledge about the target vector is needed [63]. In the second step, the positions of the objects are projected onto the output layer, where the weights are adjusted to output values (the supervised step). The trained output layer consists of $n_x \times n_y$ output neurons arranged in a squared neighbourhood. After the training, each weight of the output neurons out_j is a real number between 0.0 and 1.0. For the final prediction of classes, the response surface values must be again transformed into discrete values, zero and one. The threshold value between 0.01 and 0.99 must be determined for each class.

Consequently, the CP ANN algorithm can be explained in three steps. Firstly, a vector-represented structure of the molecule $\boldsymbol{X}_s$ is mapped into the Kohonen layer, then the weight are corrected in both the Kohonen and the output layer, and then finally the four-dimensional target – carcinogenicity – is predicted. The CP ANNs are described in the literature [31,63,64].

We took into consideration a concrete example of the chemicals in a neural network to show some aspects of the structure–activity relationship. Figure 2 demonstrates Kohonen maps for neural networks with dimension 35×35 for the training and test sets. We focused on the neuron in the position $N_x = 1$; $N_y = 8$ in the Kohonen map. Figure 3 illustrates that the structures placed in the same neuron ($N_x = 1$; $N_y = 8$) reproduce the same category of carcinogenic potency. It can be seen that these substances have a similar structure.

The molecular modelling described in this study was performed using software developed in our home laboratory, written in FORTRAN for IBM-compatible PCs and the Windows operating system. This software program *AnnToolbox for Windows* is available at the home page of the National Institute of Chemistry, Slovenia [65].

Figure 2. Kohonen maps (35×35) for training and test sets.

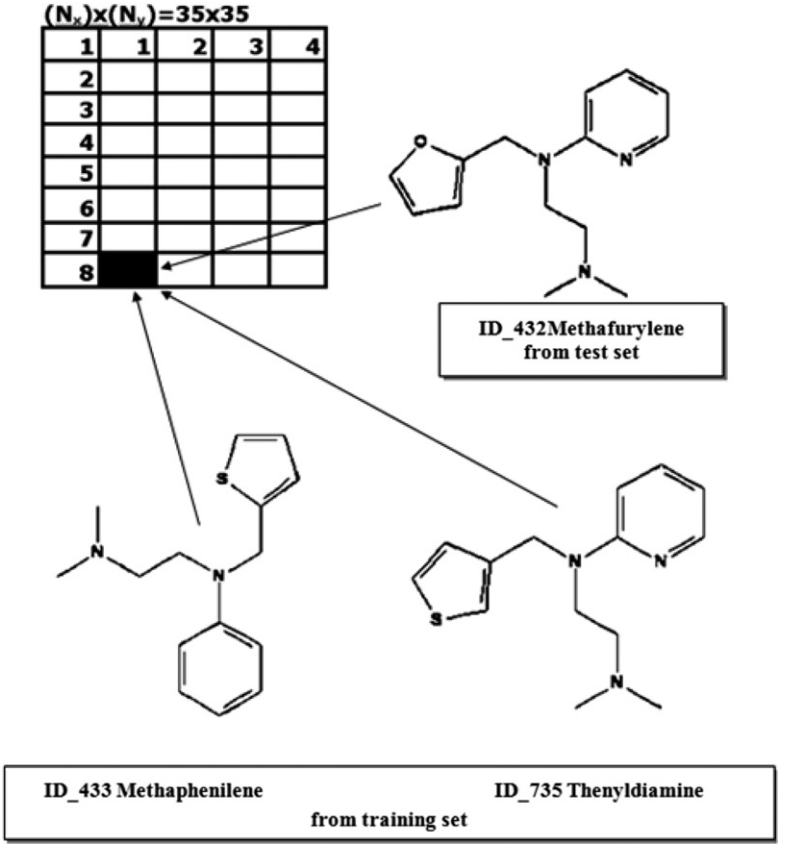

Figure 3. Neuron ($N_x = 1$; $N_y = 8$) in Kohonen map.

Table 1. Confusion matrix.

		Predicted	
		Negative	*Positive*
Observed	Negative	TN	FP
	Positive	FN	TP

Note: TP – True positive; TN – True negative; FP – False positive; FN – False negative.

A threshold (cut-off) value equal to 0.5 was applied for our best categorical models. Chemicals falling in a terminal node with mean response higher than 0.5 were classified as being positive (active or carcinogens) while chemicals falling in a terminal node with mean response lower than 0.5 were classified as being negative (inactive or non-carcinogens).

2.5 *Validation of categorical models*

The statistical performance of the models was evaluated using the following characteristics:

(1) internal performance of a training set or robustness;
(2) external performance of a test set or predictability.

A common way to evaluate the performance of the classification model (or classifier) is to employ a confusion matrix (see Table 1). The four different possible outcomes of a single prediction for a two-class problem are displayed here in a 2×2 matrix, where the rows represent the number of entries belonging to the actual class, while the columns represent the entries belonging to the predicted class. TP and TN in the Table 1 denote the number of true positives and true negatives, respectively. The number of errors made by predicting an inactive compound to be active is denoted by FP (false positives), while the number predicting an active compound to be inactive is denoted by FN (false negatives).

The statistical performance of the models has been assessed using Cooper statistics [66], which express the ability of a classification model to detect known active compounds (sensitivity), non-active compounds (specificity) and all chemicals in general (accuracy). The statistical standard binary measures used for the categorical models in the study are presented in Table 2.

The positive and negative classification rates focused more on the effects of individual chemicals, since they are conditional probabilities. Thus, the positive classification rate is a probability that a chemical classified as active is really active, while the negative classification rate gives the probability that a chemical classified as inactive is really inactive.

2.6 *Receiver operating characteristic analysis*

The receiver operating characteristic (ROC) curves are employed for a more detailed and proper analysis of classification models [67,68]. The ROC curves were first developed for signal detection [69–71]. They are substantially employed in medical tests. Recent years

Table 2. Statistical standard binary measures used for the categorical models.

Definition		*Explanation*	*Equation*
TP	True positive	Number of correct predicted 'positive'	
TN	True negative	Number of correct predicted 'negative'	
FP	False positive	Number of incorrect predicted 'positive', Type I error	
FN	False negative	Number of incorrect predicted 'negative', Type II	
ACC	Accuracy	The accuracy is the proportion of true results (both true positives and true negatives) in the population or in another words the proportion of the total number of predictions that were correct	$\mathrm{ACC} = \frac{\mathrm{TN}+\mathrm{TP}}{\mathrm{TN}+\mathrm{FN}+\mathrm{FP}+\mathrm{TP}}$
SE	Sensitivity	The proportion of positive cases that were correctly classified as positive	$\mathrm{TP\ rate} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \mathrm{Sensitivity}$
SP	Specificity	The proportion of negative cases that were correctly classified as negative	$\mathrm{TN\ rate} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = \mathrm{Specificity}$
FP rate	False positive rate	The proportion of positive cases that were incorrectly classified as negative	$\mathrm{FP\ rate} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{Specificity}$
FN rate	False negative rate	The proportion of negative cases that were incorrectly classified as positive	$\mathrm{FN\ rate} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}}$

have seen an increase of ROC graph applications in the data-mining and machine learning communities to compare different classifiers. A ROC, or ROC curve, is a graphical plot of sensitivity versus (1 – Specificity) for a binary classification system as its discrimination threshold (cut-off value) is varied. The ROC can also be represented equivalently by plotting the fraction of true positives (TPR = true positive rate; $\mathrm{TP\ rate} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \mathrm{Sensitivity}$) versus the fraction of false positives (FPR = false positive rate; $\mathrm{FP\ rate} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{Specificity}$) [66].

An ideal ROC curve would be a line along the top left-hand corner (0, 1) in ROC space, as it would not produce any false positives (or false actives). In real-world applications, this occurs only rarely. The ROC curve for a good prediction should, however, always be to the left of the diagonal between the two axes. The closer the curve tends toward (0, 1), the more accurate are the predictions made. A model with no predictive ability yields a diagonal line.

To compare two different prediction methods, both ROC curves are plotted in the same ROC space. The curve running closer to the left and top border is considered to provide a better prediction. Another good measure to compare ROC curve analysis is the

Table 3. The descriptors calculated and selected for the best models 1, 2 and 3.

Model code	*Original set of descriptors and software used for the descriptor calculations*	*Variable selection methods*	*Final sets of descriptors after their selection*
Model 1	254 descriptors generated by MDL QSAR version 2.2.2.0.7	Kohonen network and PCA	27 MDL descriptors (see Table 4)
Model 2	835 descriptors generated by DRAGON professional 5.4 (2006)	Cross-correlation matrix, multicollinearity technique, Fisher ratio and genetic algorithm	18 descriptors (12 DRAGON and 6 MDL descriptors) (see Table 5)
Model 3	88 descriptors generated by CODESSA version 2.	Cross-correlation matrix, multicollinearity technique, Fisher ratio and genetic algorithm	34 CODESSA descriptors (see Table 6)

area under the ROC curve (AUC) [72,74]. The AUC is a useful metric for an evaluation of a classifier. It is an estimator of the probability that the classifier ranks randomly chosen positive examples higher than randomly chosen negative examples. A value equal to 1.0 for a classifier indicates an optimal performance, while 0.5 indicates that the classifier performance is no better than the random method. The AUC gives an overall measure of accuracy of a predictor.

3. Results and discussion

A large number of models have been developed using the CP ANN algorithm and different sets of MDL, DRAGON and CODESSA descriptors. Methods for selection of the descriptor set have already been discussed in the Materials and method section. Finally, three sets of descriptors were employed in our study (see Tables 3–6).

A total of 805 chemicals were divided into the training and test sets as was explained previously. The CP ANN was trained from 100 to1800 learning epochs and the dimensions of the networks varied from 20×20 to 45×45 neurons. The best models correspond to a dimension of 35×35 neurons. Minimum and maximum correction factors were set to 0.01 and 0.5, correspondingly.

The main parameters of the best models 1, 2 and 3 are shown in Table 7. The statistical performance of the models is summarized in Table 8. The Cooper statistics based on the training set indicated an accuracy of 92, 91 and 93%, and a high value of the sensitivity (99, 84 and 94%) and the specificity (85, 99 and 93%) for models 1, 2 and 3, respectively. The predictive power of the models obtained was evaluated using an independent external test set. Based on this test set, the obtained accuracy was 68, 70 and 68%, with sensitivities of 73, 69 and 70% and specificities of 63, 72 and 64% for models 1, 2 and 3, correspondingly.

An important parameter of the classification models is the threshold. Figure 4 demonstrates the internal and external performance of the models depending on the threshold. In this figure the threshold is plotted versus the wrong prediction rate (FP and FN) for the training (left) and test (right) sets for models 1, 2 and 3. The prediction results

Table 4. 27 MDL descriptors selected and used for model 2.

Descriptor code	*Descriptor name*	*Definition*
MDL001	SsCH3	Sum of all (–CH_3) E-State values in molecule
MDL006	SaaCH	Sum of all (CH) E-State values in molecule
MDL007	SsssCH	Sum of all (>CH–) E-State values in molecule
MDL042	SsCH3_acnt	Count of all (–CH_3) groups in molecule
MDL052	SaasC_acnt	Count of all (CH) groups in molecule
MDL083	x0	Simple zero-order chi indices
MDL088	xp5	Simple fifth-order path chi indices
MDL105	dx0	Difference simple zero-order chi indices
MDL124	nxc3	Number of three-way clusters
MDL131	nxch7	Number of seven-membered rings
MDL148	xvpc4	Valence fourth-order path/cluster chi index
MDL160	dxvp3	Difference valence third-order path chi indices
MDL165	dxvp8	Difference valence eight-order path chi indices
MDL176	SHsOH	Sum of all [–OH] E-State values in molecule
MDL186	Hmin	Smallest atom hydrogen E-State value in molecule
MDL187	Gmin	Smallest atom E-State value in molecule
MDL193	SHarom	Sum of hydrogen E-State on aromatic CH
MDL226	LogP	Calculated value of Log *P*
MDL229	nelem	Number of chemical elements
MDL231	ncirc	Number of graph circuits
MDL235	numHBa	Number of hydrogen bond acceptors
MDL240	SHHBa	Sum of atom-type E-State indices for hydrogen bond acceptors
MDL243	Qsv	Average polarity
MDL248	sumI	Total of simple topological indices
MDL249	TTs(4) Simple	Total of valence topological indices
MDL252	totop	Total topological index based on the molecular connectivity formalism
MDL253	Wt	Total Wiener number

of the carcinogenicity are expressed as the rate of positives (active or carcinogens) and negatives (inactive or non-carcinogens). The threshold shows the difference between active and inactive compounds and thus solves the problem of separating carcinogens and non-carcinogens. A change of the threshold value from 0 to 1 leads to an increase in the prediction accuracy of non-carcinogens and a decrease in the number of false positives. In contrast, the prediction accuracy of carcinogens decreases and the number of false negative increases. This tendency is common for all our models 1, 2 and 3 no matter what set, training or test, was used.

In addition, we present the statistical performance of the models depending on the threshold of the test set. Figure 5 shows the accuracy, sensitivity (SE) and specificity (SP) against the threshold for model 1 (Figure 5A), model 2 (Figure 5B) and model 3 (Figure 5C). We have focused on maximal accuracy and plotted dotted lines to the corresponding threshold. As a result, we found the optimal threshold to be equal to 0.45 for model 1 (see Figure 5A). In this case, the accuracy has a maximal value of 0.68, the sensitivity is 0.71 and the specificity is 0.65. For model 2 (Figure 5B) the optimal threshold is 0.6 and the maximal accuracy is equal to 0.70. The sensitivity at this point is 0.69 and the specificity is 0.72. Figure 5C represents the performance of model 3. The optimal threshold

Table 5. 18 descriptors (12 DRA and 6 MDL descriptors) selected and used for model 2.

Descriptor code	*Descriptor name*	*Definition*
DRA0107	PW5	Path/walk5-Randic shape index
DRA0123	D/Dr06	Distance/detour ring index of order six
DRA0341	MATS2p	Moran autocorrelation-lag2/weighted by atomic polarizabilities
DRA0391	EEig10x	Eigenvalue 10 from edge adjacent matrix weighted by edge degree
DRA0451	ESpm11x	Spectrum moment 11 from edge adjacent matrix weighted by edge degree
DRA0464	ESpm09d	Spectrum moment 09 from edge adjacent matrix weighted by dipole moments
DRA0551	GGI2	Topological charge index of order two
DRA0565	JGI6	Mean topological charge index of order six
DRA0670	nRNNOx	Number of N-nitroso groups (aliphatic)
DRA0695	nPO4	Number of phosphates/thiophosphates
DRA0791	N-067	AI2–NH
DRA0802	N-078	Ar–N = X/X–N = X
MDL73	SdsssP_acnt	Count of all (–>P)groups in molecule
MDL113	dxp8	Difference simple eight-order path chi indices
MDL123	nxp10	Number of paths of length 10 (number of edges)
MDL159	dxv2	Difference valence second-order chi indices
MDL184	Hmax	Largest atom hydrogen E-State value in molecule
MDL229	nelem	Number of chemical elements

in this case is equal to 0.5, the maximal accuracy is 0.68, the sensitivity is 0.70 and the specificity is 0.62. A change of the threshold value leads to a revision of the sensitivity and specificity. It may be used to increase the number of correctly predicted carcinogens or non-carcinogens. After the calculations of the statistical parameters for different models, the next challenge was to find out the best model. We solved this problem using the ROC technique and by calculating the AUCs.

To compare the three different models, ROC curves were plotted in the same ROC space (see Figure 6). All three models show almost identical curves. The closer the area under the curve is to 1, the greater the predictive ability of the model. In our case, it is difficult to distinguish the differences between the models. Therefore, the areas under the ROC curves (AUCs) appeared to be more suited for comparison of the ROC curves. They were calculated for our three best models. Table 9 shows the accuracy of models 1, 2 and 3 for the training and test sets. The results are very close without a large difference. Anyway, model 2 can be estimated as the best one because the accuracy and the AUC for the test set for this model are slightly higher. Compared to models 1 and 3, model 2 has the smallest number of descriptors and learning epochs.

4. Conclusion

The CPDB rodent carcinogenic database was used for development of models for prediction of carcinogenic potency. Initial preprocessing of the data and the selection of data with carcinogenic potency for rats gave us consistent data suitable for QSAR modelling. The topological structure descriptors were calculated using the MDL,

Table 6. 34 CODESSA descriptors selected and used for model 3.

Descriptor code	*Descriptor name/Definition*
COD1	Number of atoms
COD2	Number of C atoms
COD3	Relative number of C atoms
COD4	Number of H atoms
COD5	Relative number of H atoms
COD6	Number of O atoms
COD7	Relative number of O atoms
COD8	Number of N atoms
COD9	Relative number of N atoms
COD10	Number of S atoms
COD11	Relative number of S atoms
COD14	Number of Cl atoms
COD15	Relative number of Cl atoms
COD24	Relative number of single bonds
COD25	Number of double bonds
COD26	Relative number of double bonds
COD31	Number of rings
COD32	Relative number of rings
COD33	Number of benzene rings
COD36	Relative molecular weight
COD38	Gravitation index (all pairs)
COD39	Wiener index
COD47	Kier & Hall index (order three)
COD52	Average information content (order zero)
COD54	Average structural information content (order zero)
COD60	Average information content (order one)
COD68	Average information content (order two)
COD76	Balaban index
COD77	Moment of inertia A
COD78	Moment of inertia B
COD79	Moment of inertia C
COD80	XY shadow
COD81	XY shadow/XY rectangle
COD82	YZ shadow

Table 7. The main parameters of the best CP ANN models 1, 2 and 3.

Model Code	*Descriptors*	*Dimension of network n_x/n_y*	*Number of learning epochs*	*Threshold*
Model 1	27 MDL descriptors	35×35	1000	0.45
Model 2	12 DRAGON and 6 MDL descriptors)	35×35	400	0.6
Model 3	34 CODESSA descriptors	35×35	600	0.5

DRAGON and CODESSA software programs and these provided bases for classifying molecular structures.

The CP ANN models presented in our study demonstrated good prediction statistics on the test set of 161 compounds with sensitivities of 69–73%, specificities of 63–72% as

Table 8. Statistical performance (Cooper statistics) of selected models for training and external test sets.

Model code	*Training set*			*Validation test set*		
	Sensitivity%	*Specificity%*	*Accuracy%*	*Sensitivity%*	*Specificity%*	*Accuracy%*
Model 1	99	85	92	73	63	68
Model 2	84	99	91	69	72	70
Model 3	94	93	93	70	64	68

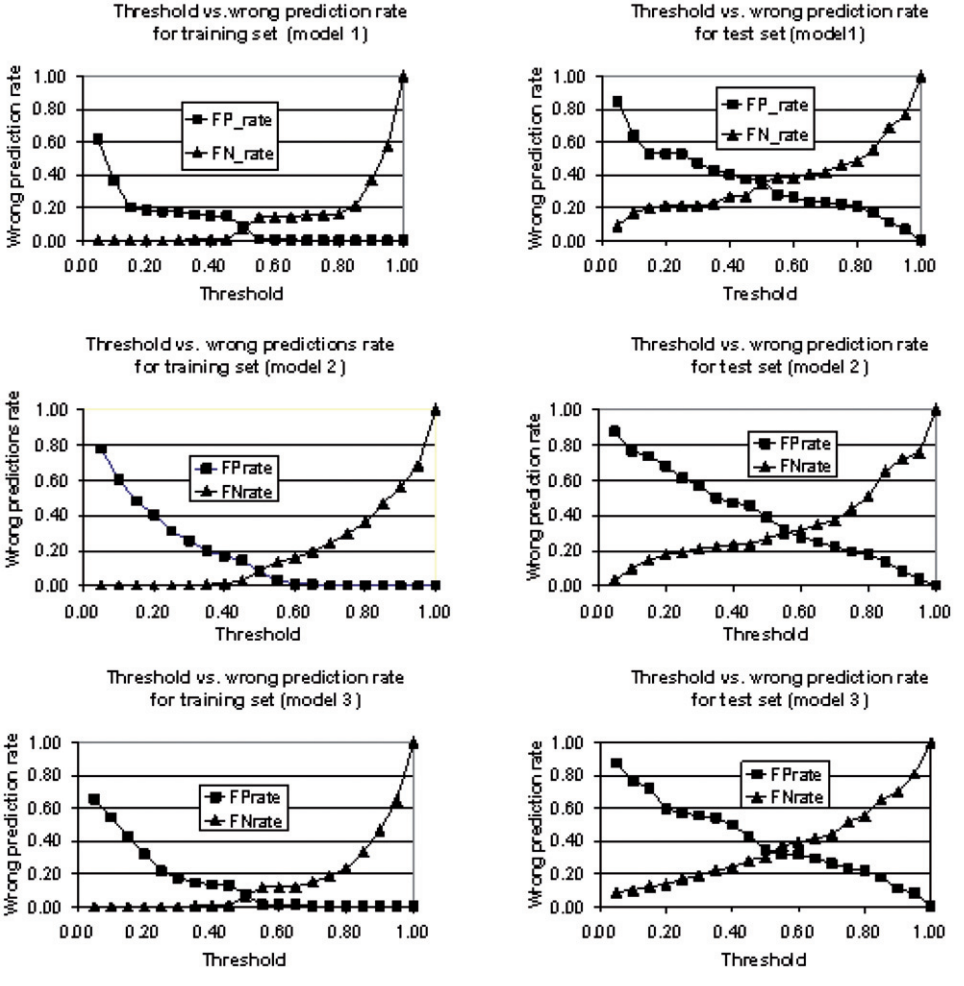

Figure 4. Threshold versus wrong prediction rate (FP and FN) for the training (left) and the test (right) sets for models 1, 2 and 3, respectively.

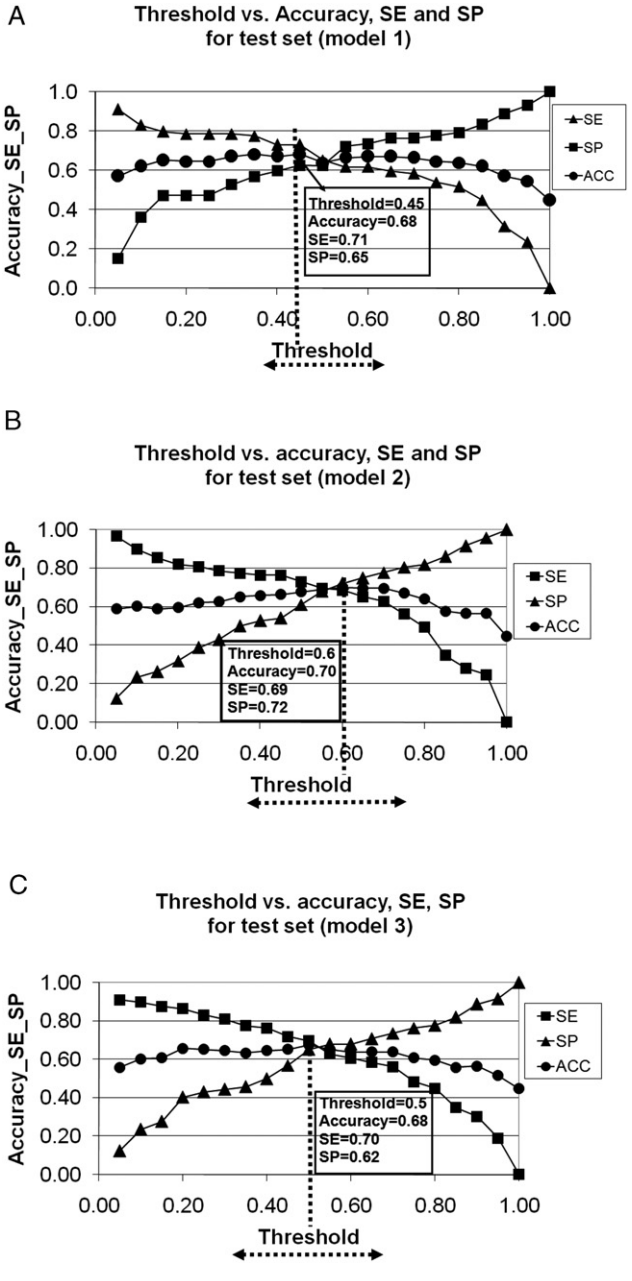

Figure 5. Threshold versus accuracy, sensitivity and specificity for the test set (A – model 1; B – model 2; C – model 3).

well as accuracies equal to 68–70%. We can vary the sensitivity and specificity of the models, changing the threshold value from 0 to 1 according to our needs and the requirements of the regulator.

Predictive toxicology programs based on the CP ANN algorithm are able to provide effective regulatory decision support information. This approach is especially useful for

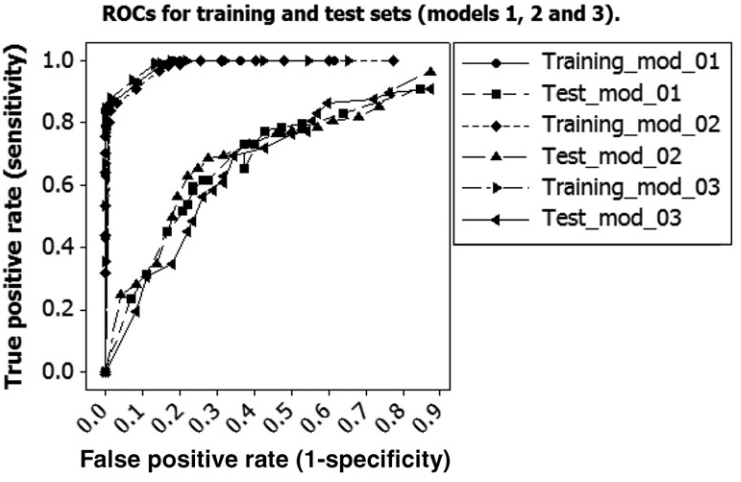

Figure 6. Receiving operation characteristics (ROCs) for the training and the test sets (models 1, 2 and 3).

Table 9. Accuracy of prediction and area under the curve (AUC) for models 1, 2 and 3.

Model Code	*Number of descriptors modelled*	*Descriptor type*	*Accuracy of training set, %*	*AUC, training set*	*Accuracy of test set, %*	*AUC, test set*
Model 1	27	MDL	92	0.988	68	0.699
Model 2	18	DRAGON and MDL	91	0.984	70	0.715
Model 3	34	CODESSA	93	0.991	68	0.680

filling data gaps in situations where toxicological data are limited. Depending on the errors in classifications, the prediction method can be used as a screening tool or as a substitute to *in vitro* and *in vivo* testing if the error is acceptable. *In silico* models can be used as a support for risk assessment for priority setting.

Acknowledgements

The financial support of the European Union through CAESAR project (SSPI-022674) as well as that of the Slovenian Ministry of Higher Education, Science and Technology (grant P1-017) is gratefully acknowledged. We would also like to thank G. Schüürmann, R. Kühne and Ralf-Uwe Ebert (Helmholtz Centre for Environmental Research, Leipzig, Germany (UFZ)) for their technical support in running the training/prediction set splitting. We would like to thank all partners of the CAESAR project for cooperation in the development of the carcinogenicity models, especially Q. Chaudry, the leader of the Central Science Laboratory (CSL DEFRA), UK, and Jane Cotterill who was involved in the project.

References

[1] N. Price, *Hail Caesar*, Chem. Ind. 15 (2008), pp. 18–19. Available at http://www.caesar-project.eu/index.php?page=results&section=results2

[2] R. Benigni and A. Giuliani, *Putting the predictive toxicology challenge into perspective: reflections on the results*, Bioinformatics 19 (2003), pp. 1194–1200.
[3] A.M. Richard and R. Benigni, *AI and SAR approaches for predicting chemical carcinogenicity: survey and status report*, SAR QSAR Environ. Res. 13 (2002), pp. 1–19.
[4] G. Patlewicz, R. Rodford, and J.D. Walker, *Quantitative structure-activity relationships for predicting mutagenicity and carcinogenicity*, Environ. Toxicol. Chem. 22 (2003), pp. 1885–1893.
[5] A.M. Helguera, M.C.A. Perez, R.D. Combes, and M.P. González, *The prediction of carcinogenicity from molecular structure*, Curr. Comput-Aided Drug Des. 1 (2005), pp. 237–255.
[6] A. Morales Helguera, M.A. Cabrera Perez, M. Perez González, R. Molina Ruiz, and H. Gonzalez-Diaz, *A topological substructural approach applied to the computational prediction of rodent carcinogenicity*, Bioorg Med. Chem. 13 (2005), pp. 2477–2488.
[7] L. Passerini, *QSARs for individual classis of chemical mutagens and carcinogens*, in *Quantitative Structure-Activity Relationship (QSARs). Models of mutagens and carcinogens*, R. Benigni, ed., CRC Press, Boca Raton, FL, 2003, pp. 81–123.
[8] R. Benigni, A. Giuliani, A. Gruska, and R. Franke, *QSARs for the mutagenicity and carcinogenicity of Aromatic Amines*, in *Quantitative Structure-Activity Relationship (QSARs). Models of mutagens and carcinogens*, R. Benigni, ed., CRC Press, Boca Raton, FL, 2003, pp. 125–144.
[9] G. Gini, M. Lorenzini, E. Benfenati, P. Grasso, and M. Bruschi, *Predictive carcinogenicity: a model for aromatic compounds, with nitrogen-containing substituents, based on molecular descriptors using an artificial neural network*, J. Chem. Inf. Comput. Sci. 39 (1999), pp. 1076–1080.
[10] G. Klopman, S.K. Chakravarti, H. Zhu, J.M. Ivanov, and R.D. Saiakhov, *ESP: A method to predict toxicity and pharmacological properties of chemicals using multiple MCASE databases*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 704–715.
[11] G. Klopman, J. Ivanov, R. Saiakhov, and S. Chakravarti, *MC4PC – An artificial intelligence approach to the discovery of quantitative structure-toxic activity relationship*, in *Predictive Toxicology*, C. Helma, ed., CRC Press, Boca Raton, FL, 2005, pp. 423–457.
[12] E.J. Matthews and J.F. Contrera, *A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASEQSAR-ES software*, Regul. Toxicol. Pharmacol. 28 (1998), pp. 242–264.
[13] Y.T. Woo and D.Y. Lai, *OncoLogic: A mechanism-based expert system for predicting the carcinogenic potential of chemicals*, in *Predictive Toxicology*, C. Helma, ed., CRC Press, Boca Raton, FL, 2005, pp. 385–413.
[14] A.A. Lagunin, J.C. Dearden, D.A. Filimonov, and V.V. Poroikov, *Computer-aided rodent carcinogenicity prediction*, Mutat. Res. 586 (2005), pp. 138–146.
[15] E. Benfenati and G. Gini, *Computational predictive programs (expert systems) in toxicology*, Toxicology 119 (1997), pp. 213–225.
[16] R. Benigni and A.M. Richard, *Quantitative structure-based modeling applied to characterization and prediction of chemical toxicity*, Methods 14 (1998), pp. 264–276.
[17] A.M. Richard, *Structure-based methods for predicting mutagenicity and carcinogenicity: are we there yet?*, Mutat. Res. 400 (1998), pp. 493–507.
[18] J.F. Contrera, E.J. Matthews, and R.D. Benz, *Prediction the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices*, Regul. Toxicol. Pharmacol. 38 (2003), pp. 243–259.
[19] G.H. Loew, M. Poulsen, E. Kirkjian, J. Ferrell, B.S. Sudhindra, and M. Rebagliati, *Computer-assisted mechanistic structure-activity studies: application to diverse classes of chemical carcinogens*, Environ. Health. Perspect. 61 (1985), pp. 69–96.
[20] E. Benfenati, R. Benigni, D.M. DeMarini, C. Helma, D. Kirkland, T.M. Martin, P. Mazzatorta, G. Ouedraogo-Arras, A.M. Richard, B. Schilter, W.G.E. Schoonen, R.D. Snyder, and C. Yang, *Predictive models for carcinogenicity: frameworks, state-of-the-art, and perspectives*, J. Environ. Sci. Health C 27 (2009), pp. 57–90.

[21] R. Benigni and C. Bossa, *Predictivity of QSAR*, J. Chem. Inf. Model. 48 (2008), pp. 971–980.
[22] R. Benigni, C. Bossa, T. Netzeva, and A. Worth, *Collection and Evaluation of (Q)SAR Models for Mutagenicity and Carcinogenicity*, European Commission Directorate General Joint Research Centre 2007, EUR 22772EN, © European Communities, 2007, pp. 1–118.
[23] D. Villemin, D. Cherqaoui, and A. Mesbah, *Predicting carcinogenicity of polycyclic aromatic hydrocarbons from back-propagation neural network*, J. Chem. Inf. Comput. Sci. 34 (1994), pp. 1288–1293.
[24] R. Benigni and A.M. Richard, *QSARS of mutagens and carcinogens: two case studies illustrating problems in the construction of models for non-congeneric chemicals*, Mutat. Res. 371 (1996), pp. 29–46.
[25] M. Vračko, *A Study of structure-carcinogenic potency relationship with artificial neural networks. The using of descriptors related to geometrical and electronic structures*, J. Chem. Inf. Comput. Sci. 37 (1997), pp. 1037–1043.
[26] S. Haykin, *Neural Networks: a Comprehensive Foundation*, Macmillan College Publishing Company, New York, 1994.
[27] J. Gasteiger and J. Zupan, *Neural networks in chemistry*, Angew. Chem. Int. Ed. 32 (1993), pp. 503–527.
[28] D.A. Winkler and D.J. Madellena, *QSAR and neural networks in life sciences*, Ser. Math. Biol. Med. 5 (1995), pp. 126–163.
[29] J. Devillers (ed.), *Neural Networks in QSAR and Drug Design*, Academic Press, London, 1996.
[30] G. Schneider and P. Wrede, *Artificial neural networks for computer-based molecular design*, Prog. Biophys. Mol. Biol. 70 (1998), pp. 175–222.
[31] J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, 2nd ed., Wiley-VCH Verlag GmbH, Weinheim, 1999.
[32] G. Schneider, *Neural networks are useful tools for drug design*, Neural Networks 13 (2000), pp. 15–16.
[33] H. Ichikawa, *Hierarchy neural networks as applied to pharmaceutical problems*, Adv. Drug. Delivery. Rev. 55 (2003), pp. 1119–1147.
[34] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 2001.
[35] G. Schneider and P. Wrede, *Artificial neural networks for computer-based molecular design*, Prog. Biophys. Mol. Biol. 70 (1998), pp. 175–222.
[36] J. Taskinen and J. Yliruusi, *Prediction of physicochemical properties based on neural network modeling*, Adv. Drug. Delivery Rev. 55 (2003), pp. 1163–1183.
[37] CAESAR project web page, 2009. Available at http://www.caesar-project.eu
[38] Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network. Available at http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html
[39] Lois Gold Carcinogenic Potency Database (CPDB). Available at http://potency.berkeley.edu/cpdb.html
[40] R. Combes, C. Grindon, M.T. Cronin, D.W. Roberts, and J.F. Garrod, *Integrated decision-tree testing strategies for mutagenicity and carcinogenicity with respect to the requirements of the EU REACH legislation*, ATLA 36 (2008), pp. 43–63.
[41] G. Schüürmann, R. Kühne, F. Kleint, R.U. Ebert, C. Rothenbacher, and P. Herth, *A software system for automatic chemical property estimation from molecular structure*, in *Quantitative Structure-Activity Relationships in Environmental Sciences*, Vol. VII, F. Chen and G. Schüürmann, eds., SETAC Press, Pensacola, FL, 1997, pp. 93–114.
[42] G. Schüürmann, R.U. Ebert, M. Nendza, J.C. Dearden, A. Paschke, and R. Kühne, *Prediction of fate-related compound properties*, in *Risk Assessment of Chemicals. An Introduction*, 2nd ed., van K. Leeuwen and T. Vermeire, eds., Springer Science, Dordrecht, NL, 2007, pp. 375–426.
[43] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Willey-VCH, New York, 2000.
[44] I. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. Palyulin, E. Radchenko, N. Zefirov, A. Makarenko, V. Tanchuk, and V. Prokopenko,

Virtual computational chemistry laboratory – design and description, J. Comput-Aided Mol. Des. 19 (2005), pp. 453–463.
[45] CODESSA PRO User's Manual, 2005. Available at http://www.codessa-pro.com/manuals/manual.html
[46] *MDL-QSAR version 2.2*, MDL Information Systems Inc., 14600 Catalina St., San Leonardo, CA 94577, USA, 2002–2004; software available at http://www.drugdiscoveryonline.com/storefronts/mdl.html
[47] DRAGON home page. Available at http://www.talete.mi.it/products/dragon_description.html
[48] M.V. Diudea (ed.), *QSPR/QSAR Studies by Molecular Descriptors*, Nova Science Publishers, Huntington, New York, 2001.
[49] M. Randić and M. Razinger, *On characterization of 3D molecular structures*, in *From Chemical Topology to Three-Dimensional Geometry*, A.T. Balaban, ed., Plenum Press, New York, 1997, pp. 159–236.
[50] A. Langewisch, F. Choobineh, and H. Deng-Yuan, *Selection procedures in linear models*, J. Stat. Plan. Infer. 54 (1996), pp. 271–277.
[51] A.D. Walmsley, *Improved variable selection procedure for multivariate linear regression*, Anal. Chim. Acta 354 (1997), pp. 225–232.
[52] Y. Kano, *Variable selection for structural models*, J. Stat. Plan. Infer. 108 (2002), pp. 173–187.
[53] Y.S. Shih and H.W. Tsai, *Variable selection bias in regression trees with constant fits*, Comput. Stat. Data. Anal. 45 (2004), pp. 595–607.
[54] B. Lucić, N. Trinajstić, S. Sild, M. Karelson, and A.R. Katritzky, *A new efficient approach for variable selection based on multiregression: prediction of gas chromatographic retention times and response factors*, J. Chem. Inf. Comput. Sci. 39 (1999), pp. 610–621.
[55] C.L. Waller and M.P. Bradley, *Development and validation of a novel variable selection technique with application to multidimensional quantitative structure-activity relationships studies*, J. Chem. Comput. Sci. 39 (1999), pp. 345–355.
[56] W. Zheng and A. Tropsha, *Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle*, J. Chem. Comput. Sci. 40 (2000), pp. 185–194.
[57] K. Tang and T. Li, *Comparison of different partial least-squares methods in quantitative structure-activity relationships*, Anal. Chim. Acta 476 (2003), pp. 85–92.
[58] L. Xu and W.J. Zhang, *Comparison of different methods for variable selection*, Anal. Chim. Acta 446 (2001), pp. 475–481.
[59] F. Despagne and D.L. Massart, *Variable selection for neural networks in multivariate calibration*, Chemom. Intell. Lab. Syst. 40 (1998), pp. 145–163.
[60] G. Castellano and A.M. Fanelli, *Variable selection using neural-network models*, Neurocomputing 31 (2000), pp. 1–13.
[61] Z. Ramadan, X.H. Song, P.K. Hopke, M.J. Johnson, and K.M. Scow, *Variable selection in classification of environmental soil samples for partial least square and neural network models*, Anal. Chim. Acta 44 (2001), pp. 231–242.
[62] A. Yasri and D. Hartsough, *Toward an optimal procedure for variable selection and QSAR model building*, J. Chem. Comput. Sci. 45 (2001), pp. 1218–1227.
[63] J. Zupan, M. Novič, and I. Ruisanchez, *Kohonen and counterpropagation artificial neural networks in analytical chemistry*, Chemom. Intell. Lab. Syst. 38 (1997), pp. 1–23.
[64] P. Mazzatorta, M. Vračko, A. Jezierska, and E. Benfenati, *Modeling toxicity by using supervised Kohonen neural networks*, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 485–492.
[65] *AnnToolbox for Windows*, National Institute of Chemistry, Ljubljana, Slovenia; software available at http://www.ki.si/en/display-pages/equipment/?tx_ukki_pi1%5Buid%5D=318&cHash=e267f7b447
[66] J.A. Cooper, R. Saracci, and P. Cole, *Describing the validity of carcinogen screening test*, Br. J. Cancer. 39 (1979), pp. 87–89.
[67] T. Fawcett, *An introduction to ROC analysis*, Pattern Recognit. Lett. 27 (2006), pp. 861–874.

[68] J. Fan, S. Upadhye, and A. Worster, *Understanding receiver operating characteristic (ROC) curves*, Can. J. Emerg. Med. 8 (2006), pp. 19–20.
[69] H.L. Van Trees, *Detection, Estimation, and Modulation Theory (Part I)*, Wiley, New York, 1968.
[70] J.P. Egan, *Signal Detection Theory and ROC Analysis*, Academic Press, New York, 1975.
[71] J.A. Swets, *Measuring the accuracy of diagnostic systems*, Science 240 (1988), pp. 1285–1293.
[72] A.P. Bradley, *The use of the area under the ROC curve in the evaluation of machine learning algorithms*, Pattern Recognit. 30 (1997), pp. 1145–1159.
[73] J.A. Swets, *Signal Detection Theory and Roc Analysis in Psychology and Diagnostics*, *Collected Papers*, Lawrence Erlbaum Associates, Mahwah, NJ, 1996.
[74] F. Provost and T. Fawcett, *Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions*, Proceedings of the Third International Conference of Knowledge Discovery and Data Mining (KDD-97), AAAI Press, Newport Beach, CA, 1997, pp. 43–48.