NATIONAL INSTITUTE OF CHEMISTRY, LJUBLJANA, SLOVENIA

Carcinogenicity QSARs models using non-cogeneric chemicals for regulatory purposes.

Natalja Fjodorova, Marjana Novich, Marjan Vrachko, Marjan Tushar
Laboratory of Chemometrics, National Institute of Chemistry, SI-1000 Ljubljana, Slovenia
natalja.fjodorova@ki.si

Abstract In the context of EU legislation, such as REACH and the Cosmetics Directive (Council Directive 2003/15/EC), it is anticipated that (Q)SARs will be used more extensively, in the interests of time- and cost-effectiveness and animal welfare.

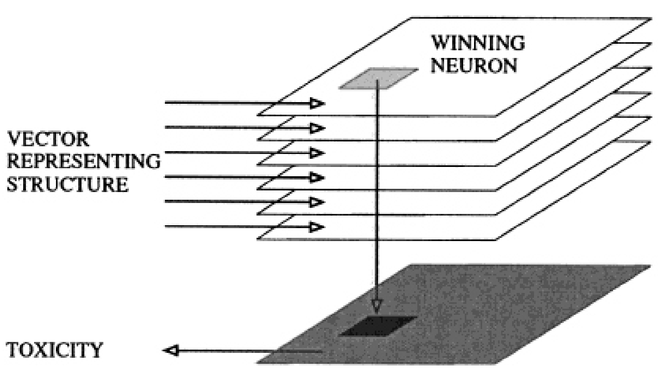
A survey and analysis of QSARs models for carcinogens for cogeneric classes of chemicals has shown good statistical performance (70-100%correct prediction). However such local models are limited in number by lack of sufficient data. Therefore models for non congeneric chemicals have been developed in scope of European Commission (EC) funded project CAESAR in accordance with principals of validation adopted by Organization for Economic Cooperation and Development (OECD).

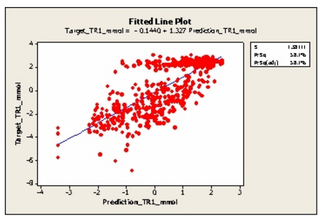
In silico models for prediction of the ability of chemicals to induce carcinogenicity in rodent using Counter Propagation Artificial Neural Network (CP ANN) have been built and analysed. Statistical performance of models have been discussed.

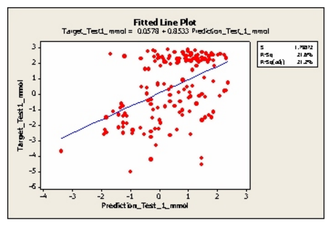
Data: The analyzed dataset consist of **805 chemicals** extracted from Carcinogenic Potency Database (CPDBAS). Original data table with **1481chemicals** has been taken from **Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network** http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html. The molecular structure were represented as MDLmolfiles. The molecular structure information was obtained as topological structure descriptors, including atom-type and group-type E-state and hydrogen E-state indices, molecular connectivity, chi indices, topological polarity, and counts of molecular features. The MDL QSAR software computed all these descriptors.

Method: Counter-Propagation Artificial Neural Network (CP-ANN).

The architecture of CP ANN is presented in Figure 1. Basically, it is built up from two layers of neurons arranged in two-dimensional rectangular matrix. The input or Kohonen layer contains information on input values (descriptors) while the output layer is associated to output values (logTD50 in quantitative models and 0 or1 values in classification model). The learning procedure is different in both layers. In the input layer the learning is the same as in Kohonen network. It means that after the learning the objects are organized in such a way that similar objects are situated close to each other. It is to emphasize that only the input values participate in this phase of learning (unsupervised step). In the second step the positions of objects are projected to the output layer, where the weights are adjusted to output values (supervised step).







Two plots on the left presentes the correlation between target and prediction for one of our intermediate model for training (at the top) and for test set (at the bottom).

Figure1. Architecture of the CP ANN.

Descriptors selection and minimization

Initial dataset contained 254 MDL descriptors for 805 chemicals (644molecules in training set and 161 molecules in test set)

Step1: 94MDL descriptors selection using Kohonen map.

Step2: 86MDL descriptors selection eliminating zero and constant values.

Step3: 27MDLdescriptors selection using Principle Component Analysis (PCA).

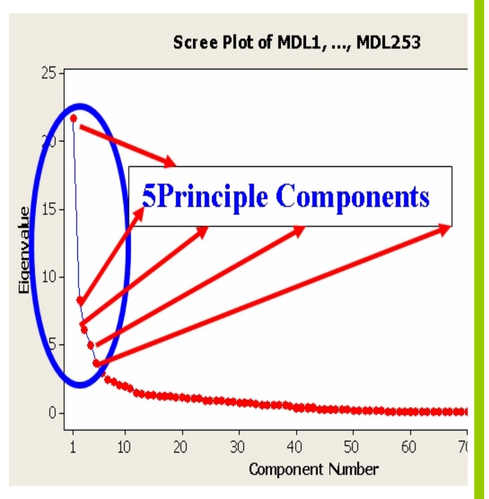
Kohonen map: Kohonen neural network of dimension 7X7 was applied, which enables one to map objects into 49 positions. Similar objects were mapped into the same position (x,y coordinate of the Kohonen map). Kohonen network was trained until a limiting error is reached. 1-2 descriptors from each neurons were chosen on the basis of smallest and largest distance between the neuron and descriptorsvector.

Fragment of Kohonen map 4X4 is presented in the Table2.

Table2. Fragment of Kohonen map 4X4 with selected descriptors.

Ny/nx	1	2	3	4
1	245 TTs(4) Simple MDL249 248 totop MDL252	110 dxp7 MDL112 112 dxp9 MDL114	128 nxch6 MDL130 150 xvch6 MDL152	132 nxch10 MDL134 154 xvch10 MDL156
2	117 nxp6 MDL119 124 nxpc4 MDL126	86 xp5 MDL088 226 nrings MDL230	190 si MDL194 194 idwbar MDL198	185 SHCsatu MDL189
3	139 xvp6 MDL141 146 xvpc4 MDL148	122 nxc3 MDL124 144 xvc3 MDL146	81 x0 MDL083 244 W MDL244	1 SsCH3 MDL001 41 SsCH3_acnt MDL042
4	28 SssO MDL028 227 ncirc MDL231	195 idc MDL199 240 W MDL244	220 ka2 MDL224 228 phia MDL232	218 k3 MDL222 221 ka3 MDL225

Table3. Principle Cmponent Analysis.



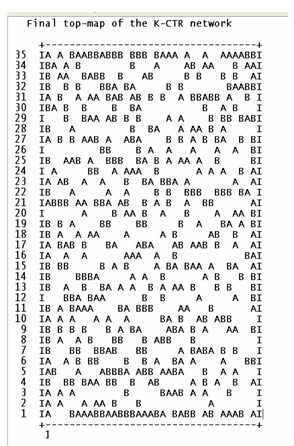
Principle Component Analysis. : We have used Principle Component Analysis to form a smaller number of uncorrelated variables and to avoid multicollinearity in dataset of descriptors. 5Principle components acount 52% of total data variation.

Kohonen maps for Trainting/Test set

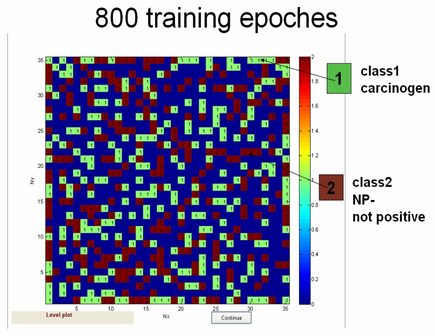
Classification models

	Training		Test	
	Accuracy	Sens./Spec.	Accuracy	Sens./Spec.
CPANN_model 27MDL descriptors	0.99	0.99/ 0.98	0.64	0.70/ 0.58
Two layer classifier combined method (IRFM)	0.57	0.77/ 0.36	0.63	0.79/ 0.44

Intermediate results



Kohonen map for model 35*35 and 800 training epoches



Continuouse data models (Quantitative models)

Models	Reduction of descriptors method, model	TRAINING		TEST	
		R_train	RMSE	R_test	RMSE
CP ANN_model 250MDLdescriptors		0.74	1.51	0.47	1.78
CP ANN_model 86MDLdescriptors	Kohonen map	0.72	1.54	0.42	1.90
CP ANN_model 27MDLdescriptors	PCA	0.74	1.52	0.45	1.80
SVM_model (Thomas Ferrary) 86MDLdescriptors		0.82	1.23	0.47	1.81

Not satisfied results for quantitative models

Asknowleagement
The financial support of the European Union through CAESAR project (SSPI-022674) is gratefully acknowledged.

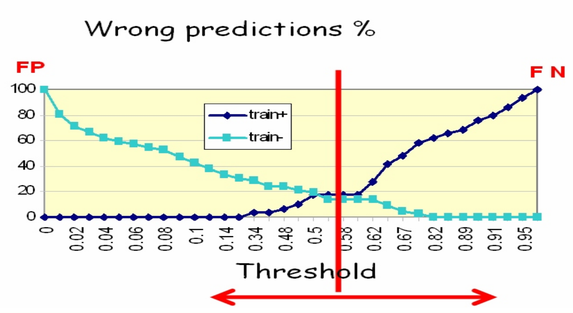
Results interpretation

Confusion matrix 2classes (Positive- Negative)

Training set / Test set	Positive (P)	Negative (N)	Conflict	N =644 / N =161
Train/ Test (P)	TP 244/48	FN 51/ 28	78/ 19	327/ 95
Train/ Test (N)	FP 0/ 21	TN 266/ 39	51/ 6	317/ 66

Accuracy-(AC),
true positive(TP), true negative (TN)
AC=(TN+TP)/(TN+TP+FN+FP)
TPrate=Sensitivity=TP/(TP+FN)
TNrate=Specificity=TN/(TN+FP)

Classification model
Reduce False Negative Predictions



From regulatory perspective, the higher sensitivity in predicting carcinogens is more desirable than high specificity.

Reduce false negative prediction can be done using threashold adjastment as shown in the Classification

Current efforts: improvement of models

- Optimization of CP ANN.
- Integration of different models (genotox).
- Implementation of structure alert (SA) approach like **toxtree** for separation of compounds with or without genotoxic carcinogenicity alerts, and without carcinogenic activity alerts.
- To focus model to high **sensitivity** in prediction of carcinogenicity potency.