WORKSHOP ON QSAR MODELS FOR REACH
Mario Negri Institute, Milan, Italy - March 10-11, 2009

Natalja Fjodorova, Marjana Novic, Marjan Vracko

Kemijski institut Ljubljana Slovenia
Ljubljana, Slovenia

The model for carcinogenicity

Overview

- Carcinogenicity (identification, evaluation and classification criteria)
- Carcinogenic potency prediction modeling:

- data used for modeling;

- steps in modeling:

Splitting dataset into training and test sets;

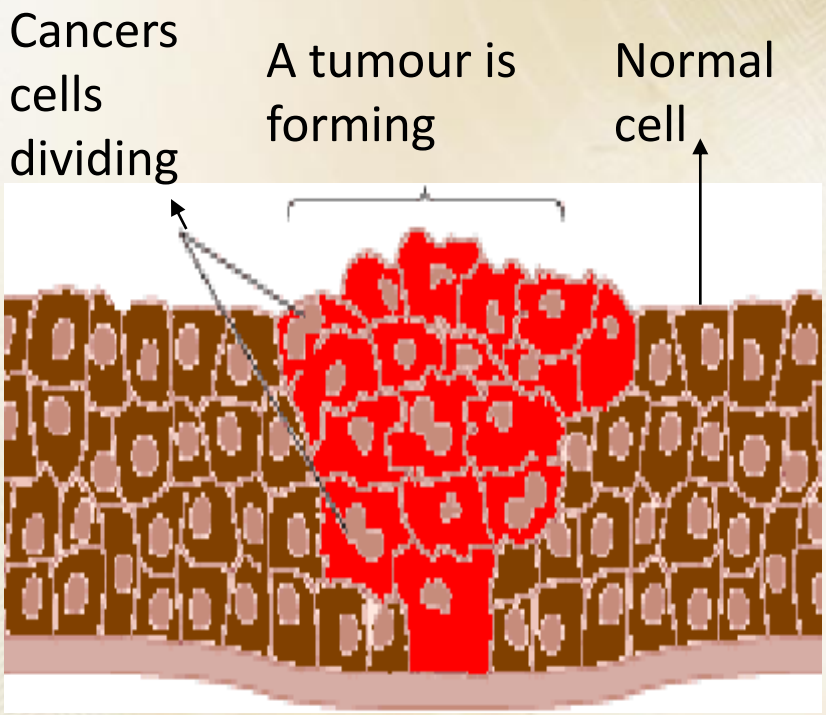
Calculation and selection of descriptors;

Applied algorithms;

- Statistical performance of obtained models and their evaluation

Conclusions

Carcinogenicity



The term “carcinogen” generally refers to an agent, mixture, or exposure that increases the age-specific incidence of cancer.

Carcinogen identification is an activity grounded in the evaluation of the results of scientific research.

How do we evaluate evidence of cancer?

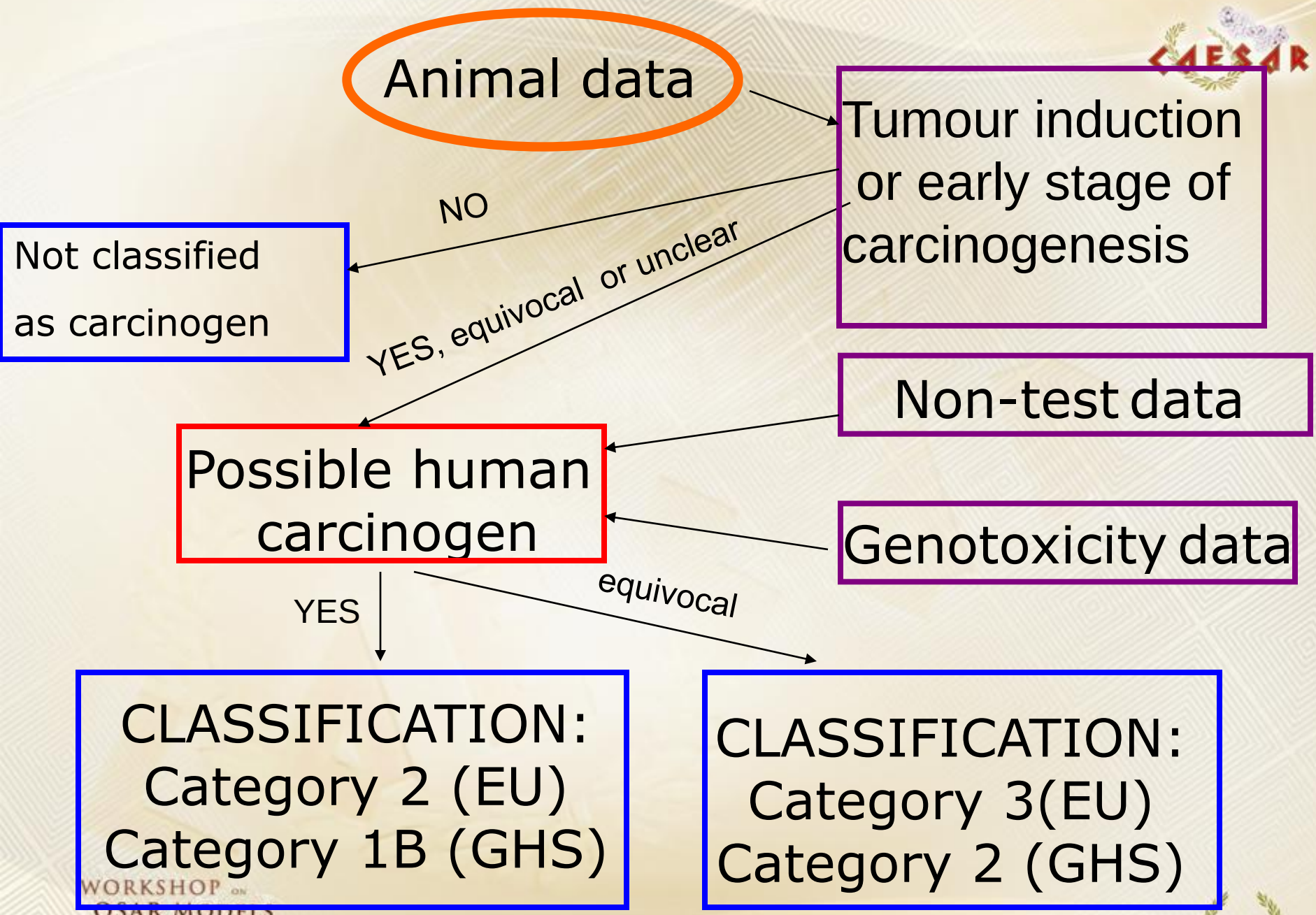
- 1.Studies of carcinogenicity in humans
- 2.Carcinogenicity studies in animals
- 3.Other relevant data
- additional evidence related to possible carcinogenicity
  - Genetic Toxicology
  - Structure-Activity Comparisons
  - Pharmacokinetics and Metabolism
  - Pathology

Each source of data has a role in the overall assessment.

Cancer Risk Assessment

IARC International Agency for Research of Cancer

	IARC		For animals
Group	**Classification**	**Explanation**	**Classification**
Group A	Human Carcinogen	sufficient human evidence for causal association between exposure and cancer	
Group B1	Probable Human	limited evidence in human	
Group B2	Probable Human	inadequate evidence in humans and sufficient evidence in animals	**clear evidence**
Group C	Possible Human Carcinogen	limited evidence in animals	**some evidence**
Group D	Not Classifiable as Human Carcinogenicity	inadequate evidence in animals	**equivocal**
Group E	No Evidence of Carcinogenicity in Human	at least two adequate animal tests or both negative epidemiology and animal studies	**no evidence**


Animal data
Tumour induction or early stage of carcinogenesis
NO
Not classified as carcinogen
YES, equivocal or unclear
Non-test data
Possible human carcinogen
Genotoxicity data
YES
equivocal
CLASSIFICATION: Category 2 (EU) Category 1B (GHS)
CLASSIFICATION: Category 3(EU) Category 2 (GHS)

- The chemicals involved in the study belong to different chemical classes,

(non congeneric substances)

- The work addresses industrial chemicals, referring to the REACH initiative. The aim is to cover as much chemical space as possible

According to the OECD principles QSAR models should:

(1) Be associated with a defined endpoint of regulatory importance;

(2) Take the form of an unambiguous algorithm;

(3) Have a defined domain of applicability;

(4) Be assosiated with appropriate measures of goodness-of-fit, robustness and predictivity

(5) Have a mechanistic interpretation, if possible.

http://appli1.oecd.org/olis/2007doc.nsf/linkto/env-jm-mono(2007)2

Principle 1- **A defined endpoint**

Endpoint is the property or biological activity determined in experimental protocol, (OECD Test Guideline).

Carcinogenicity is a defined endpoint addressed by an officially recognized test method (**Method B.32** Carcinogenicity test – **Annex V** to Directive **67/548/EEC**).

Carcinogenic potency of chemicals in the rodent bioassay:

1. Yes/NO response (if a chemical has to be considered carcinogenic or not in various experimental groups)
2. A carcinogenic potency index TD50 is the dose tolerated by half of the animals to remain tumourless for each induced tumour type.
3. The profile of tumours (e.g.target organs) induced by the chemical.

Dataset:

805 chemicals were extracted

from rodent carcinogenicity study findings for

1481 chemicals

taken from the Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network

http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html

derived from the Lois Gold Carcinogenic Database (CPDBAS)

What was done

to ensure quality and consistency of data

1. We focused only on well defined organic compounds therefore e.g. mixtures, polymers, inorganic compounds, metallorganic compounds, salts, complexes and compounds without well defined structure were excluded;

2. Only data for rats were used as data for rats are more close to human;

3. Cross-checking of structures by at least two partners of the consortium.

Three errors were found in the structures (acknowledged in the EPA website) and one in the toxicity value.

CAESAR Classification ranges

Classes	TD50, mg/kg_bw/day	Total compounds	Training Set	Test Set
Carcinogen	TD50 ≤ 3000	421	332	89
No carcinogen	TD50 > 3000 (NP)	384	312	72
		805	644	161

Principle 2- **An unambiguous algorithm**

- The algorithm is the form of relationship between the chemical structure and property or biological activity being modelled.
- Examples:

1. Statistical (regression) based QSARs
2. Neural network models, which include both learning processes and prediction processes.

Transparency in the (Q)SAR algorithm can be provided by means of the following information:

a) Details of the training/test sets used to develop the algorithm.

b) Definitions of all descriptors in the algorithm, and a description of their derivation

c) Definition of the mathematical form of a QSAR model, or of the decision rule (e.g. in the case of an SAR)

Splitting dataset into training/test sets

805 chemicals

(421 carcinogens and 384 non-carcinogens)

were split into

training set (644 chemicals) and

test set (161 chemicals)

(This work was performed by UFZ Centre for Environmental Research– (Germany))

Descriptors calculated for modeling:

254 MDL descriptors calculated by MDL QSAR software,

835 Dragon descriptors calculated by DRAGON software,

88 CODESSA descriptors calculated using CODESSA software

Selection of descriptors:

- The goal was to establish a reasonable number of predictor variables to ensure a good generalized performance and to reduce data "noise".

- A lot of different approaches were applied.

- The best results were obtained using a hybrid selection algorithm (HSA), which combines the genetic algorithm (GA) concepts and a stepwise regression.

F Ros, M Pintore, JR Chretien (2002) Molecular description selection combining genetic algorithms and fuzzy logic: application to database mining procedures, Chemom. Intell. Lab. Syst. 63, 15-26

Eight MDL descriptors were selected using a hybrid selection algorithm for the best models

MDL_ID	Index	Definition	Descriptors categories
MDL005	SdsCH	Sum of all (= CH –) E-State values in molecule	Atom-Type E-State
MDL051	SdssC_acnt	Count of all (= C <) groups in molecule	Atom-Type E-State Acnt
MDL062	SdsN_acnt	Count of all (= N –)groups in molecule	Atom-Type E-State Acnt
MDL114	dxp9	Difference simple 9th order path chi indices	Connectivities simple
MDL130	nxch6	Number of 6-membered rings	Connectivities subgraph counts
MDL187	Gmin	Smallest atom E-State value in molecule	HE-State Categories
MDL190	SHCsats	sum of hydrogen E-State on sp3 C on saturated bond	HE-State for Groups
MDL210	SHBint2_Acnt	Count of internal hydrogen bonds with 2 skeletal bonds between donor and acceptor	Internal H-bonds E-State

Methods used by partners to develop models

- Adaptive Fuzzy Partition- (AFP);
- Counter-Propagation Artificial Neural Network- (CP-ANN);
- K Nearest Neighbour- (KNN);
- Self-organising Networks of Active Neurons based on the Group Method of Data Handling- (GMDH);
- Combined models.

Partner	NIC-LJU	BCX	CSL	KM
SAR/ QSAR	QSAR	SAR	QSAR	(Q)SAR
Descriptor software	MDL; DRAGON; CODESSA	MDL	MDL; DRAGON; CODESSA	NIC_models KM_models CSL_models
Modelling Method	CPANN	AFP	KNN (k=3) (ADMEWORKS Modelbuilder)	GMDH CO-NN (cost-benefit matrix: (0; 30; 500; -200))

KM- (KnowledgeMiner Software Germany) generated QSAR models that are optimal with respect to both prediction power of a model and an a priori given cost-benefit matrix along with the uncertainty level of the experimental toxicity values.

One of the best models was obtained using CP ANN method, therefore we focused on this method here.

NIC_LJU applied Neural Networks as an algorithm for modeling

What are the advantages and disadvantages of this method?

Neural Networks

Advantages

1. Capable of modeling multivariate data with non-linear functions.
2. Easier to use than traditional nonlinear statistical methods**.** Neural networks *learn by example*.
3. Prediction accuracy is generally high.
4. Output may be expressed as discrete or continuous values (response surface).
5. Fast prediction of the target values.

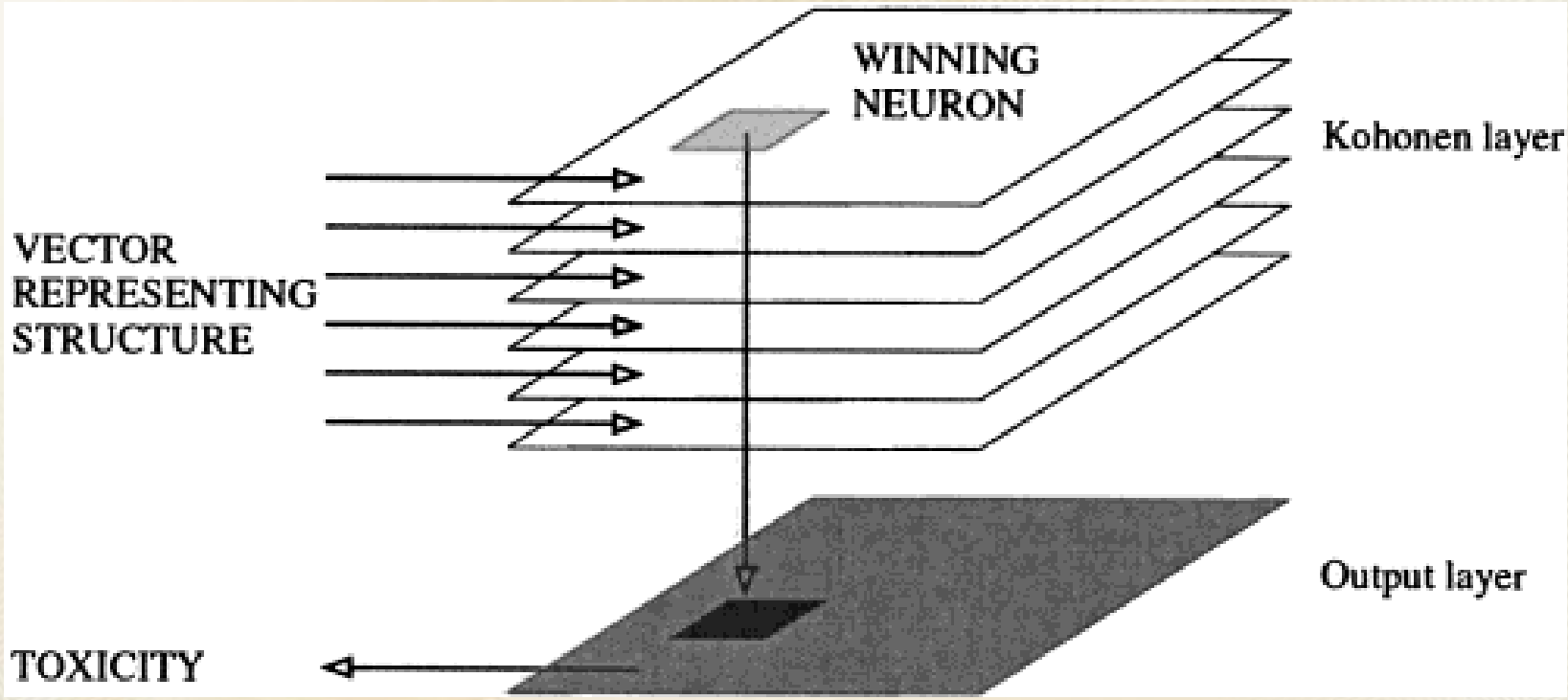
Disadvantages

1. Not possible to extrapolate outside the boundaries of the training set.

Counter Propagation Artificial Neural Network

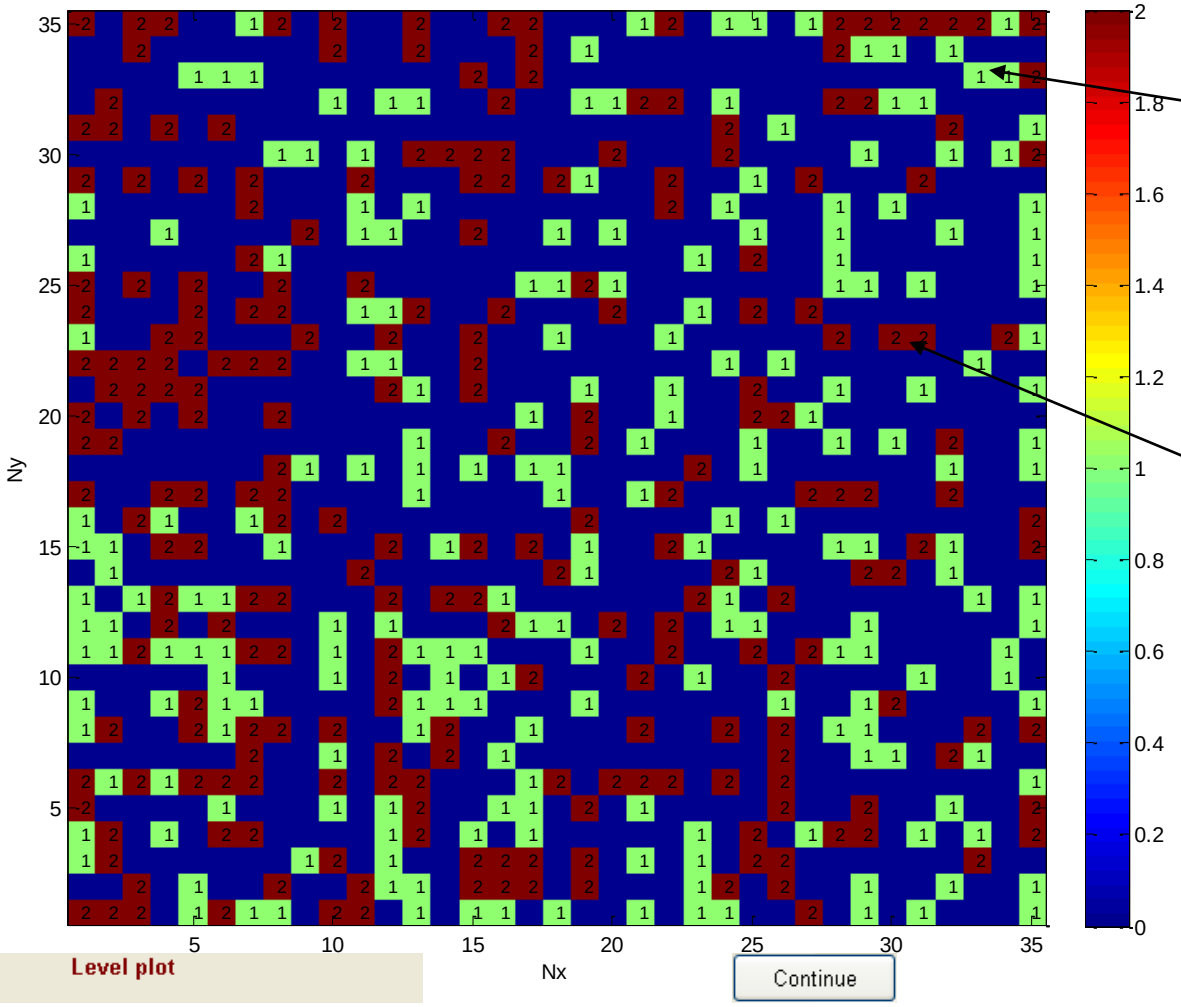
Step1: mapping of molecule Xs (vector representing structure) into the Kohonen layer

Step2: correction of weights in both the Kohonen and the Output layer



Step3: prediction of target (toxicity)
Ts=carcinogenicity

Output layer-Kohonen map for model (nxn=35*35 and 800 training epochs)

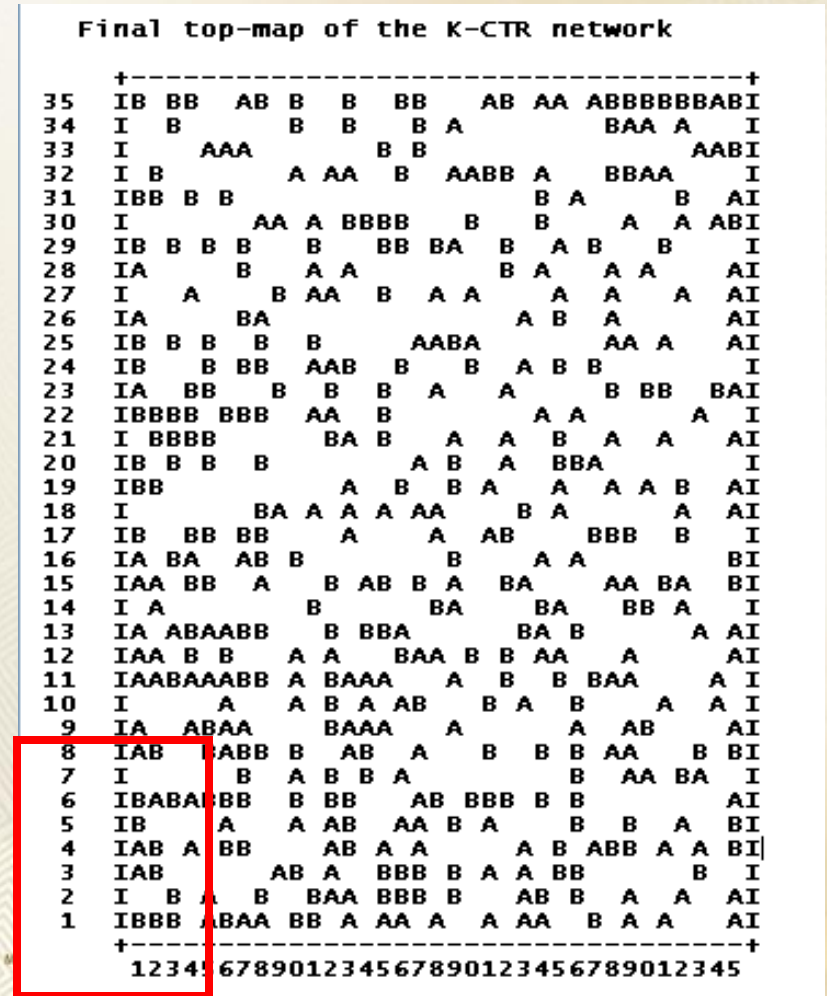


+ active-carcinogen

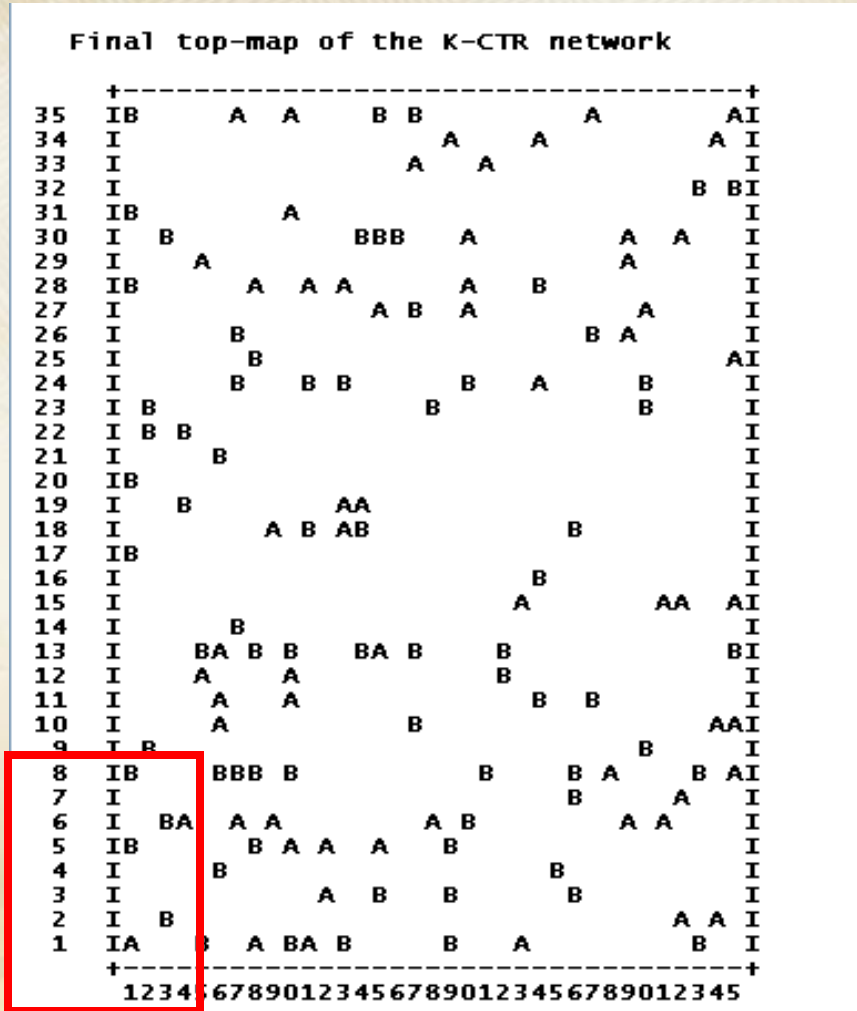
- non active-non carcinogen

Kohonen maps (35x35) for training and test sets

Training set



Test set



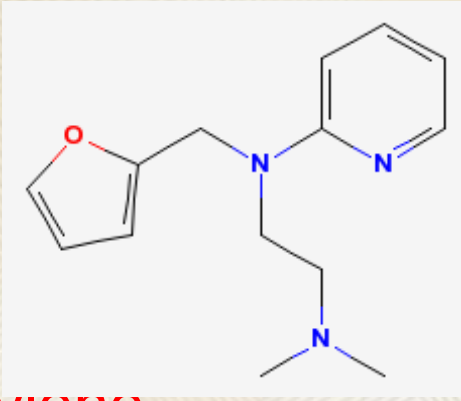
Neuron (Nx=1;Ny= 8) in Kohonen map

Structures placed in the same neuron reproduce the same value of toxicity or carcinogenicity

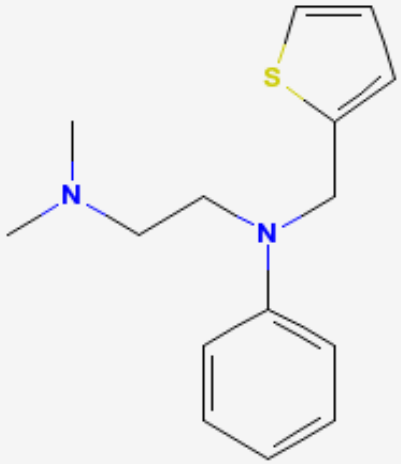
$(N_x)x(N_y)=35x35$

1	1	2	3	4
2				
3				
4				
5				
6				
7				
8				

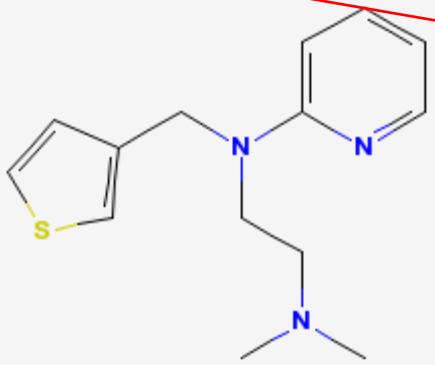
From test set



ID_432 Methafuryiene



ID_433 Methaphenilene



ID_735 Thenyldiamine

From training set

Development of a CP ANN model

Different parameters have been used to identify the model.

1. Number of neurons in x and y direction-

20x20; 25x25; 30x30; 35x35; 40x40;

2. Number of learning epochs-

100, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800;

3. Different sets of descriptors (MDL, DRAGON, CODESSA)

Final model

The final algorithm uses fixed parameters

1. Number of neurons in x and y direction-
   35x35
2. Number of learning epochs-
   800
3. 8 MDL descriptors

Principle 3- A Defined Domain of Applicability

The definition of the Applicability Domain (AD) is based on the assumption that a model is capable of making reliable predictions only within the structural, physicochemical and response space that is known from its training set.

- List of basic structures (for example, aniline, fluorene..)
- The range of chemical descriptor values.

Domain of applicability for the model with 8 MDL descriptors

MDL_ID	Min_value of descriptor	Max_ value of descriptor
MDL_005	0.000	30.74
MDL_051	0.000	18.000
MDL_062	0.000	4.000
MDL_114	-0.4009	7.7061
MDL_130	0.000	7.000
MDL_187	-5.0185	2.000
MDL_190	0.000	66.2633
MDL_210	0.000	8.000

Predictive Toxicology Approaches

1. **Classification or qualitative models**

Response- YES/NO principle

YES- P-positive or active or carcinogen

NO- NP-not positive or not active or non carcinogen

Results from

NIC_LJU (Slovenia), BCX (France) and CSL (England) models

2. Quantitative models (QSARs) Continuous data prediction on the basis of experimental evidence of rodent carcinogenic potential

Response- TD50_Rat- Carcinogenic potency in rat (expressed in mmol/kg body wt/day) –

Results from

Istituto di Ricerche Farmacologiche "Mario Negri" (IRFMN, Italy)

Principle 4- **Appropriate measures**

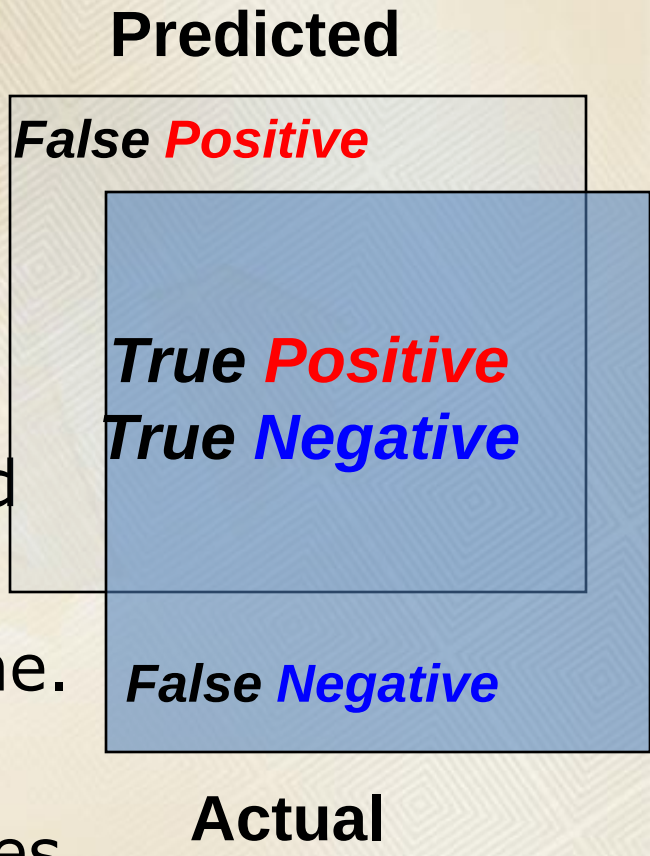
- **goodness-of-fit,**
- **robustness** (internal performance) and
- **predictivity** (external performance)

What is known about statistical performance of the model?

The assessment of model performance is sometimes called statistical validation.

Evaluation of the Classification System

- **Training set** represents class values for learning.
- **Test set** represents class values for evaluation
- **Evaluation:** Hypotheses are used to establish classification in the test set, which is compared to known one.



- **Accuracy:** percentage of examples in the test set that are classified correctly.

Confusion matrix for two classes

		Predicted	
		Negative	Positive
Observed	Negative	TN	FP
	Positive	FN	TP

True positive **(TP)** True negative **(TN)**

False positive **(FP)** False negative **(FN)**

Accuracy (**AC**) =(TN+TP)/(TN+TP+FN+FP)

Sensitivity**(SE)**=TP/(TP+FN)

Specificity**(SP)**=TN/(TN+FP)

Eleven best classification models

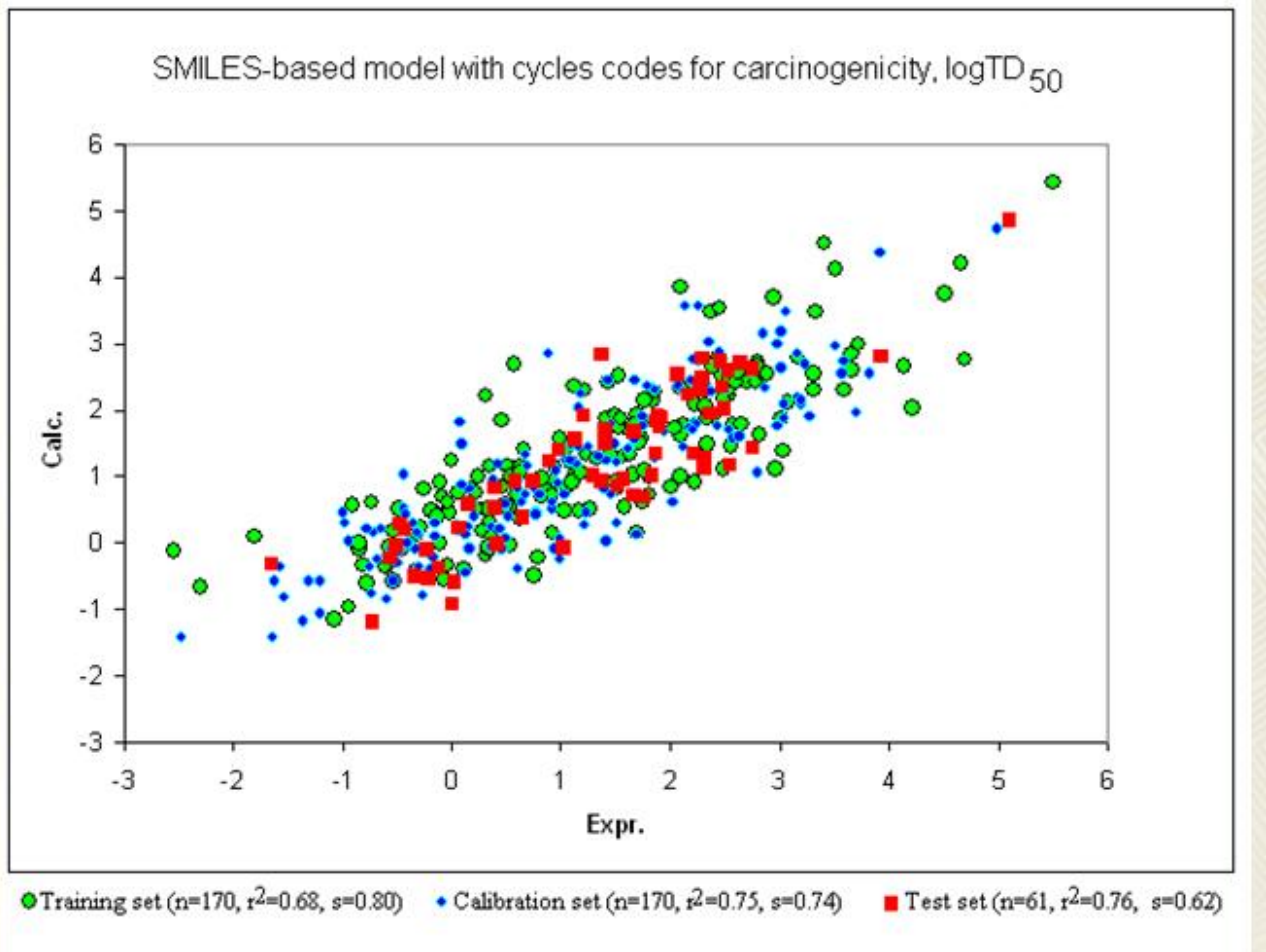
Model code	A	B	C	D	E	F	G	H	I	K	L
Partner	NIC-LJU	BCX	CSL	CSL	NIC-LJU	NIC-LJU	NIC-LJU	NIC-LJU	CSL	CSL	CSL
Descriptor type	MDL	MDL	Dragon MDL	Codessa	MDL	Co-dessa	Dragon MDL	MDL	Dragon; MDL	MDL	Dragon Co-dessa MDL
Number of Descrip-tors	8	8	18	38	27	38	18	27	18	14	34
Modelling method	CP ANN	AFP	KNN	KNN	CP ANN	CP ANN	CP ANN	CP ANN	KNN	KNN	KNN

Validation statistics derived from the best models A and B using 8 MDL descriptors

	Model A CP ANN method		Model B AFP method	
	Training	Test	Training	Test
Accuracy, %	91	73	71	70
Cross-validation, %	66		60	
Sensitivity (Carcinogen), %	96	75	73	72
Specificity (Non-Carcinogen), %	86	69	69	68

Model A	Training set	Test set
Total compounds (number)	644	161
Accuracy,%	91	73
Cross-validation (leave 20% out), %	66	
False Positive (FP) (number)	44	22
False Positive Rate, %	14	31
False Negative (FN) (number)	13	22
False Negative Rate, %	4	25
Postive Predictive Value (PPV) (precision), %	88	75
Negative Predictive Value (NPV), %	95	69
Sensitivity (Carcinogen), %	96	75
Specificity (Non-Carcinogen), %	86	69

Quantitative models developed in collaboration with ChemPredict Project



Training set
(170 chemicals)
R^2= 0.68

Calibration set
(170 chemicals)
R^2= 0.75

Test set
(61 chemicals)
R^2= 0.76

Conclusions

Classification or qualitative models prediction power:

Accuracy of the training set is 0.91- 0.96;

Accuracy of the test set is 0.68- 0.74,

Sensitivity is 0.69- 0.75;

Specificity 0.63- 0.72.

Quantitative models (QSARs) prediction power:

R^2 for test set = 0.76.

CAESAR's models

can be used as support for carcinogenicity assessment, both in classification and with potency evaluation, for instance to evaluate relative risk of different compounds, or of metabolite or parent compound.

Acknowledgements

The financial support of the European Union through CAESAR and ChemPredict projects is gratefully acknowledged

CAESAR partners:

IRFMN; CSL; BCX; POLIMI; KM; LJMU; UFZ; TNO

THANK YOU!