QSTR Modelling of Acute toxicities on fathead minnow (Pimephales promelas) ***by counterpropagation neural networks***

Pierluigi Barbieri[*+], Nadege Piclin[o], Andrzej Szymoszek[*], Mariana Novic[*], Marian Vracko[*], Emilio Benfenati[o]

[*]Kemijski inštitut Ljubljana, Hajdrihova 19, 1001 Ljubljana, p.p. 3430, Slovenija

[o] Istituto di Ricerche Farmacologiche "Mario Negri", Via Eritrea 62, 20157 Milano, Italy

[+]on leave from Dip. Scienze Chimiche, Universita' di Trieste, Via Giorgieri 1, 34127 Trieste, Italy

Introduction

- Risk assessment of chemicals requires evaluation of:
  - Exposure to chemicals, their dispersion in in the environment (environmental monitoring)
  - Hazard/toxicity of chemicals towards humans and/or living beings in the environment.

- Experimental studies to determine hazard of chemicals are expensive and up to now performed only for a small number of chemical compounds (few thousands in comparison with the millions registered in the CAS registry) so:

- Quantitative Structure Activity Relationships studies for modelling toxicities (QSTR) can help.

Introduction: experimental

The toxicity towards Fathead Minnow (*Pimephales promelas*) – a freshwater fish from north America - has been tested [1] for

•562 compounds representing a cross section of industrial organic chemicals [2], and

•Toxicity has been reported as median lethal concentrations LC50 (mmol/l) after 96 hours exposure

1. C.L. Russom, S.P. Brandbury, S.J. Broderius, D.E. Hammermeister, D.A. Drummond, *Environmental Toxicology and Chemistry*, 16 (1997) 948-967.

2. G.D. Veith, B. Greenwood, R.S. Hunter, G.I. Niemi, R. Regal, *Chemosphere*, 17 (1988) 1617-1630 .

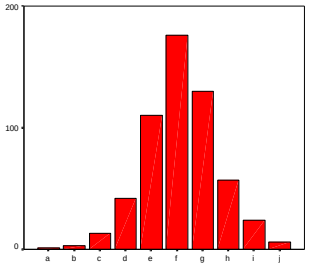
Target values: the toxicities

In order to obtain an index directly proportional to the toxicity of the chemicals, 96 h LC50s were transformed into

$$Tox = (\log10(1/LC50)).$$

Classes of *Tox* can be identifyed by rounding the values of *Tox*, round(*Tox*) ranging from -3 to 6.

round(*Tox*)	Frequency	%	Cumulative %
6	1	0.18	0.18
5	3	0.53	0.71
4	13	2.31	3.02
3	42	7.47	10.5
2	110	19.57	30.07
1	176	31.32	61.39
0	130	23.13	84.52
-1	57	10.14	94.66
-2	24	4.27	98.93
-3	6	1.07	100
Total	562	100	



562 chemical structures

•562 compounds: different chemical classes, aromatic, non aromatic, heteroatoms,

•MW 32:488 AUs.

- rather wide base for building up a model with some ambition of **generality**,
- need of checking for **subgroups** to be modelled separately and for **ouliers**.

Our approach follows the framework of **Quantitative Structure Toxicity Relationships (QSTR).**

•The chemical structures can be represented as **molecular descriptors** which are numbers extracted by a well defined algorithm from a representation of the molecules.

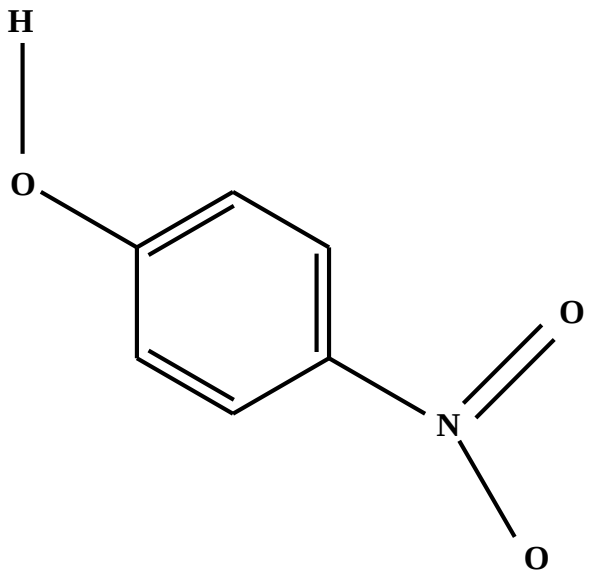
159 Topological, physicochemical and electronic descriptors based on 2D graphs [3] were considered:

- Size descriptors (molecular volume, molecular weight and van der Waals volume);
- Shape indices which account for the ramification degree, the oblong character, etc.;
- Connectivity indices describing degree of branching and cyclization in the compounds;
- Information contents descriptors;
- Lipophilicity descriptor, Electronegativity descriptor;
- Electrotopological State indices (electron density at each atom or hybrid group);

Experimental log $P_{oct/water}$ has been considered as a descriptor as well.

3. SciQSAR, SciVision, 200 Wheleer Rd, Burlington, Ma 101803 USA

Representations of a molecule



Topological picture of p-nitrophenol

P-nitrophenol.mol

```
     15 15  0  0  0                 1 V2000
        0.7133    0.7255   -0.0126 C   0  0  0  0  0
        0.7437   -0.7117   -0.0096 C   0  0  0  0  0
       -0.4565   -1.3849   -0.0126 C   0  0  0  0  0
       -1.6987   -0.6617   -0.0137 C   0  0  0  0  0
       -1.7265    0.7450   -0.0094 C   0  0  0  0  0
       -0.5009    1.4165   -0.0114 C   0  0  0  0  0
       -2.9010    1.4308   -0.0058 O   0  0  0  0  0
        2.0267   -1.4274   -0.0036 N   0  3  0  0  0
        2.0719   -2.5700   -0.0065 O   0  0  0  0  0
        3.2884   -0.6963    0.0154 O   0  5  0  0  0
        1.6532    1.3000   -0.0146 H   0  0  0  0  0
       -0.4854   -2.4862   -0.0118 H   0  0  0  0  0
       -2.6490   -1.2200   -0.0152 H   0  0  0  0  0
       -0.4932    2.5189   -0.0115 H   0  0  0  0  0
       -2.6902    2.3790   -0.0064 H   0  0  0  0  0
      1  2  2  0  0  0
      1  6  1  0  0  0
      1 11  1  0  0  0
      2  3  1  0  0  0
      2  8  1  0  0  0
      3  4  2  0  0  0
      3 12  1  0  0  0
      4  5  1  0  0  0
      4 13  1  0  0  0
      5  6  2  0  0  0
      5  7  1  0  0  0
      6 14  1  0  0  0
      7 15  1  0  0  0
      8  9  2  0  0  0
      8 10  1  0  0  0
    M  END
```

Exploratory data analysis: Kohonen NN

Each of the 562 molecules is described by a vector of 160 descriptors and for it a toxicity value has been measured.

The computational tool used for **exploratory data analysis** is Artificial Neural Network known as Self Organizing Map (SOM) or as **Kohonen neural network** [4].

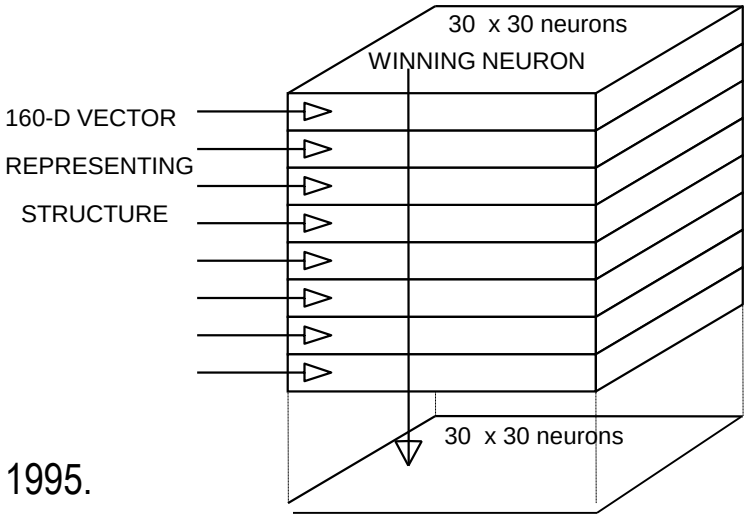
The Kohonen neural network is a rectangular array of neurons (I.e."prototype" vectors – for us dim. 160-chosen to represent the input data). In our case study we decided to use a square (30x30) network.

In the beginning (***initialization***), "prototype" vectors approximate poorly the data.

Then we present all objects/molecules to this network and start the learning (**training phase**).

This is a non-linear algorithm.

At the end of training, the neurons are arranged in such a way that the similar neurons are close to each other.



4. T. Kohonen "Self Organizing Maps", Springer Verlag, 1995.

Aim: to **recognize compounds** presenting structures **very different from the others**, thus being hard to model.

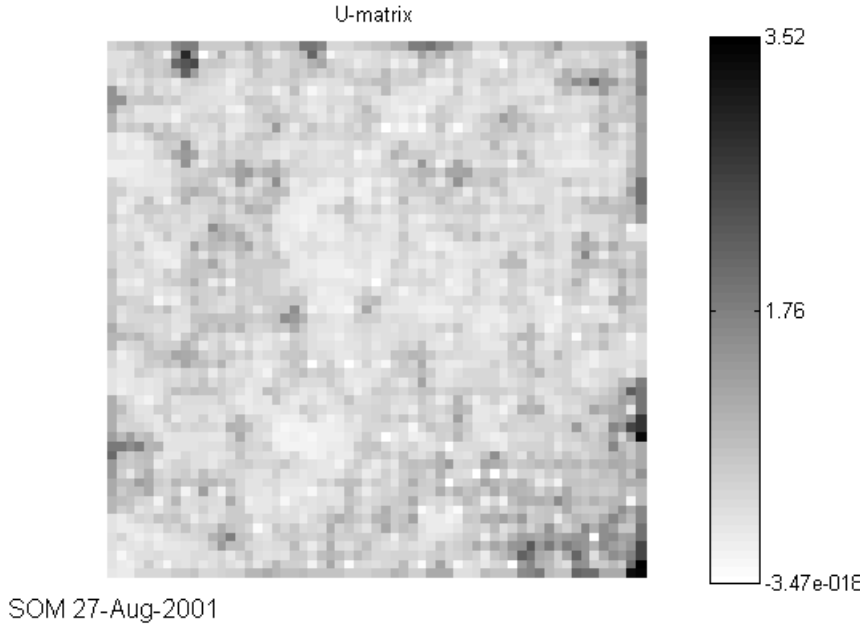
Outliers detection by SOM has been studied in [5]: two types of outliers are known:

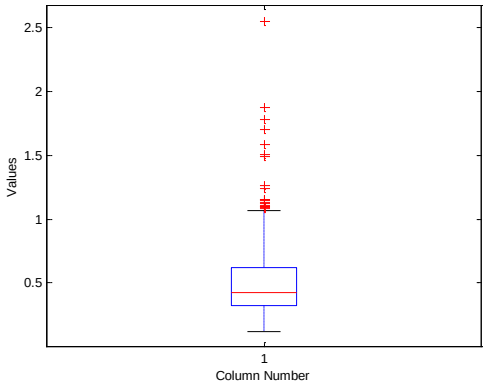
Type 1) compounds having as "**best matching neuron**" an "**outlying neuron**", that is a "prototype" vector very different from other "prototype" vectors (degree of similarity can be measured as euclidean distance);

Type 2) compounds having big Quantization Errors: the "distance" between the compound and its best matching neuron is very high.

Graphic tool that shows heterogenehity of neurons is **unified distance matrix**: it gives a measure of dissimilarity between neurons in 160-D input space

$\Rightarrow$ Identify outlying neurons (outliers type1) .

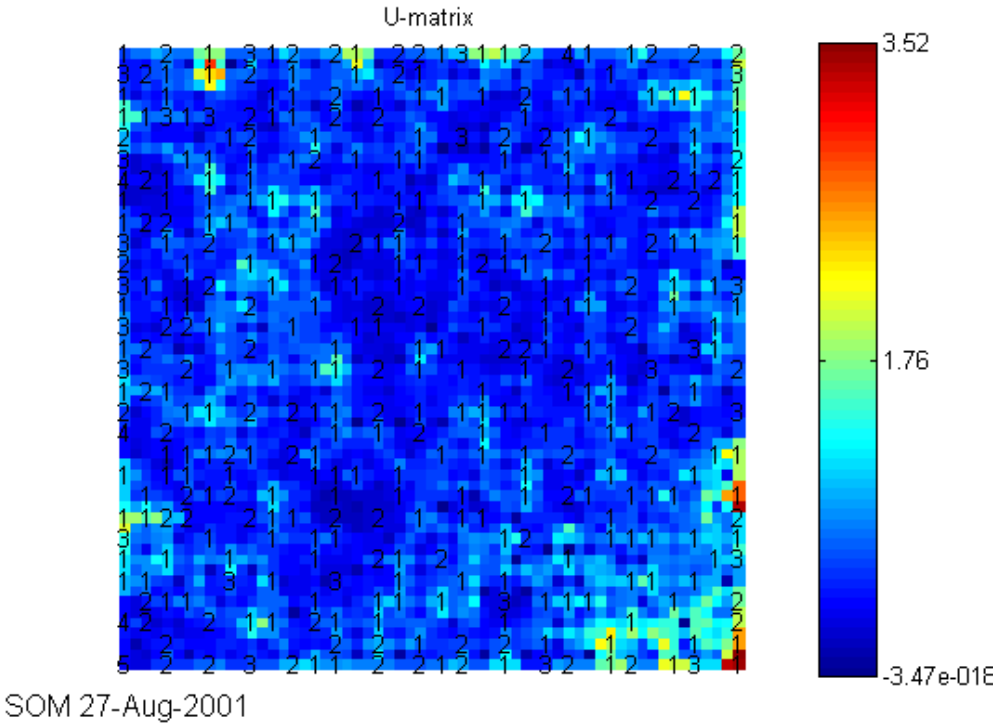




Quantization errors (outliers type2) can be highlighted by boxplots.

5. A. Muñoz, J. Muruzábal, Neurocomputing, 18 (1998) 33-60

Distribution of 562 chemicals on the Kohonen network; colorbar indicates degree of similarity between neurons. →



- Some outliers (type I) are highlighted;
- **Data have been splitted in** two sets:
  - the **training set** for building the model (371compounds=66%of 562),
  - the **test set** for evaluating the performance of the model on unknown c. (191 c.=33%),

The following **precautions** were adopted so that the neural network for predicting toxicities will be trained and will learn from structures as much heterogeneous as possible:

- the map has been gridded into 25 submaps and 66% of compounds were taken in **each** submap;
- outlying compounds (I and II type) were put into the training set;
- most toxic compounds were put into the training set.

Modelling: counter propagation neural network

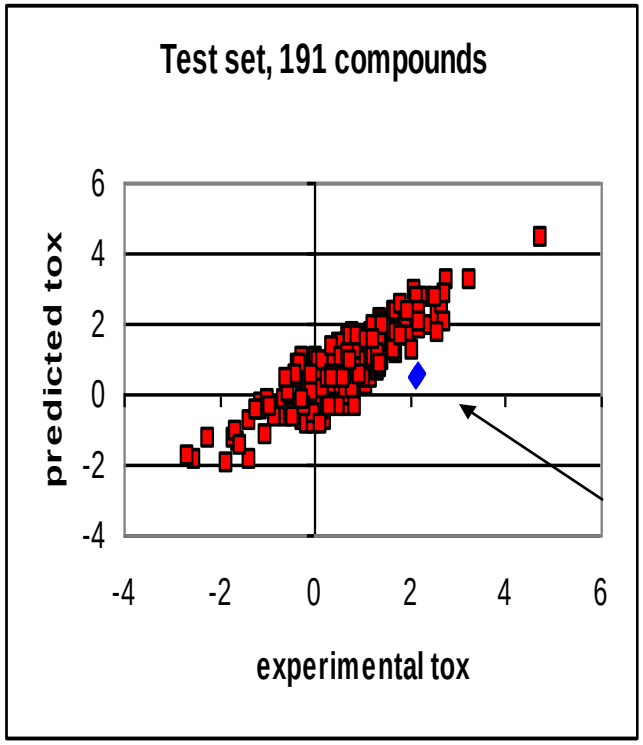
- **Counter propagation neural network** [6] provide an architecture that allows to connect Kohonen layers accounting for the molecular structures to a layer of neurons accounting for the toxicities of molecules;
- During training (i.e. iterative presentation of the 371 input data to the network) the network learns how to relate structural information to toxicity values.
- This Quantitative Structure-Toxicity Relationship is a model.

- If a network is trained very long, it models very well the training data, but probably it models also the noise associated to experimental values (toxicities) and it looses generality (performs badly with unknown/test data). This is the problem of **overtraining**.
- In our case Networks have been trained with different number of epochs for learning (1000, 2000,…,5000), in order to verify if longer or shorter learning produces network giving better prediction. The network trained for 2000 epochs performs best.

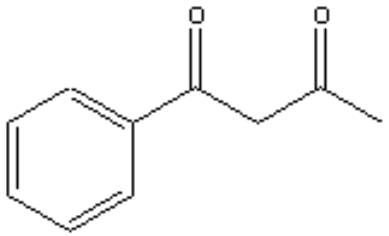
training of cp-nn (2000 epochs)	*b0*	*b1*	*r*	R^2
training set (371 compounds)	0.036	0.961	0.983	0.97
test set (191 compounds)	0.117	0.869	0.877	***0.77***

6. J. Zupan, J. Gasteiger " Neural Networks in Chemistry and Drug Design", John Wiley, 1999.

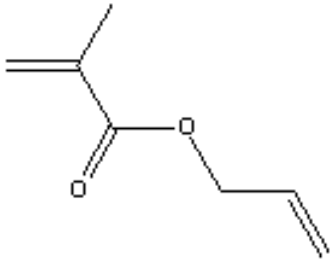
- The plot of predicted vs target values for the test set in figure shows that few compounds are badly modelled, so that relatively low predictions corresponds to high experimental values.



The two compounds badly modelled are: Benzoylacetone



Allyl methacrylate



Which descriptors are relevant? Hunting correlations...

- **Planes of the network representing descriptors and toxicity can be displayed [7], and they can be ordered in a matrix so to display planes of highly correlated descriptors close to each other [8].**

- **In order to show correlations between descriptors, the weights of neurons (the layers of the counterpropagation network) are considered as variables: a Kohonen map is built for:**

abs(cov(weights))

7. J.Himberg, J.Ahola, E.Alhoniemi, J.Vesanto and O.Simula "The Self-Organizing Map as a Tool in Knowledge Engineering", in Nikhil R. (Editor), "Pattern Recognition in Soft Computing Paradigm", World Scientific Publishing, (2001) 38-65.

8. J.Vesanto, J.Ahola, "Hunting for Correlations in Data Using the Self-Organizing Map", in H.Bothe, E.Oja, E.Massad, C.Haefke (Editors),"Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications", ICSC Academic Press (1999), 279-285

Hunting correlations...

A “second order” Kohonen map displaying 160 descriptors + *Tox:*

Hmax		numHBd	SHsOH	k0	SHBint2		24-SssNH	SHssNH	NHBint6	NHBint5	38 - SsF		28 - SdsN	19-SssssC	15 - StsC	14 - SddC	10 - StCH
Hmaxpos		34 - SsOH			Redundancy		NHBint2			SHBint3			13-SsssCH	26 - StN	52-SdssS	NHBint10	48 - SsSH
				NHBint4						dxvp3	dxvp4	dxvp5	xvch8		xvch7	SHBint5	53-SdddssS
		SHBint4			MAX(ES)		31-SddsN	dxv2		dxvp7	dxvp6		xvch5	NHBint7	xvch9	NHBint8	48 - SsSH
Qsv										dxvp8					xvch3	53-SdddssS	SHtCH
Qv					SHCsats					dxvp10	dxvp9		Wt	knotpv		30 - SsssN	SHBint9
			9-SssCH2		7 - SsCH3			dx0					dxp10			8 - SdCH2	NHBint9
	Hmin										dxp9	xvp10			50-SssS		76 - SsI
		LOG(POW)				k3				dxp8		xvch10		16 - SdssC			75-SssssSn
VCE						ka3				xp10		xvp9				NHBint3	SHBint6
				nelem				46-SdsssP		dxp7					49 - SdS		
	MMES	SMVE			Gmin		36 - SssO			xp9		xvp8	xvp5	SHCsatu			SHtvin
												xvp7	xvp6				21-SsNH2
numHBa				dxv1		dx1			dxp6	xvch6				xvpc4			SHsNH2
	35 - SdO										xp8			xvp4			
Gmax			dxv0			totop			xp7	tets2			xvp3				55 - SsCl
								dxp5			idw						SHCHnX
sumdell				knotp		dxp4	nrings				W		xv2				
									xp6						MES		VES
si				dxp3							idc						phia
nclass		dx2						xp5					xv1				ka2
					SHvin	11-SdsCH									Qs		k2
			Wp														
sumI	xpc4		xp4					SHother		idcbar		ka1					LgP
				TTs(4) Simp							k1			xv0			
TTd(4) Val					k0				x1								
xp3	Pf	x2			Info content		idwbar		nvx		x0		fw				Tox

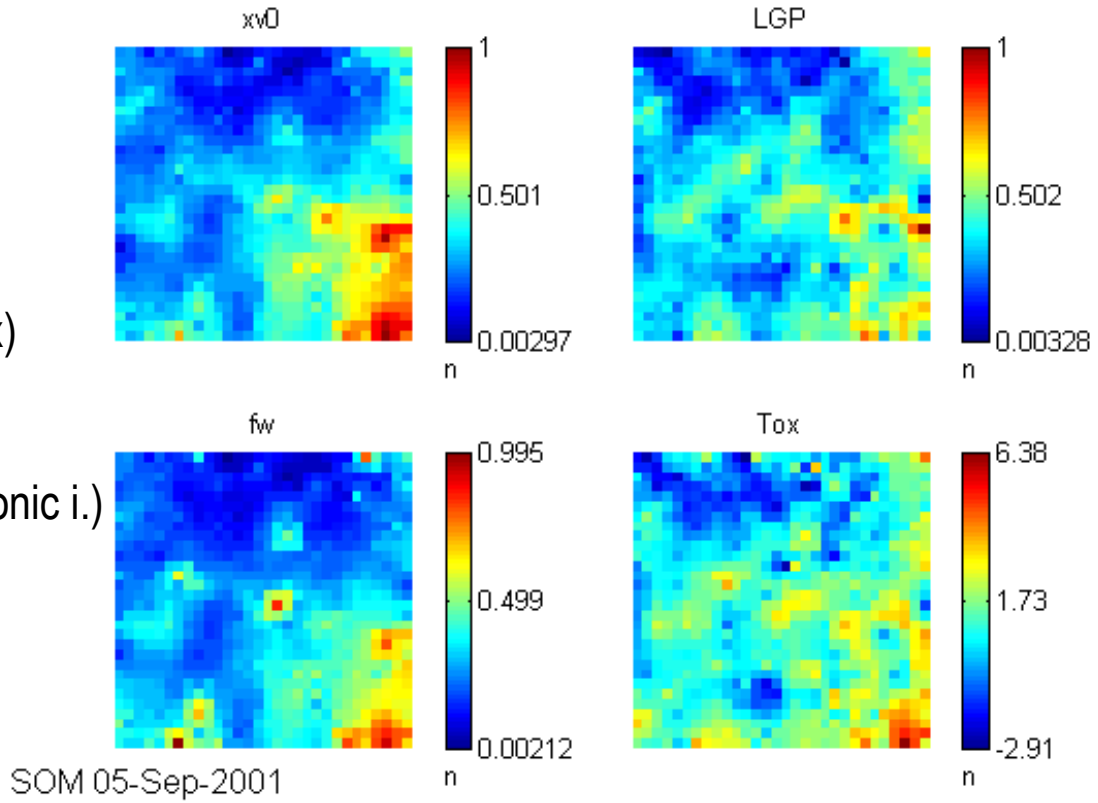
The descriptors being more similar to Tox...

… are

LGP=Experimental LogP (lipophylicity index)

fw = weight of the free compound (steric i.)

xv0 = valence index of Kier and Hall (electronic i.)



This is in agreement with the approach of Hansch [9], according to whom biological responses depends additively on lipophylic, steric and electronic properties of molecules.

9. C.Hansh, A.J. Leo “Substituents constants for correlation analysis in chemistry and biology”, Wiley, New York, 1979

Conclusions

- A **MU**ltistep **M**odelling **P**rocedure based on **S**elf-**O**rganization has been outlined and applied to find relationships between descriptors of chemical structures and toxicity on the fish *Pimephales promelas*.
- Methods for outliers detections were implemented so to improve the split of the original data in training and test set.
- Quantitative methods for finding (local) correlations between descriptors, up to now applied only to Kohonen neural networks, have been applied to counter propagation NN.
- The test on 191 compounds not used to train the model shows that the model is robust (R^2_{test}=0.77), and that 2D descriptors performs well.

Acknowledgement

The investigations have been accomplished within the framework of the Research Training Network IMAGETOX (Intelligent Modelling Algorithms for the General Evaluation of TOXicities);

The European Commission is acknowledged for supporting financially the network (HPRN-CT-1999-00015).

THANKS!