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

Qasim Chaudhry, Marco Pintore

BioChemics Consulting S.A.S., France

The CAEAR Model for Developmental Toxicity

Developmental Toxicity

- ***Developmental toxicity*** has been defined as "adverse effects induced during pregnancy, or as a result of parental exposure," that "can be manifested at any point in the life span of the organism" (UNECE, 2004).

- Cost for each experiment: in the range of many 100,000's euros

Data set – Molecular structures

- Extracted from Arena *et al.* (2004) including 293 cpds
- Structural quality check: remaining **292 cpds**

  - **Checking Names, structures, CAS etc** by online databases: ChemFinder (http://chemfinder.cambridgesoft.com), ChemIDPlus (http://chem.sis.nlm.nih.gov/chemidplus/);

  - **Searching duplicate chemicals and isomers**;

  - **Removing ions and neutralizing molecules;**

  - **Cross-checking** by at least 2 different partners.

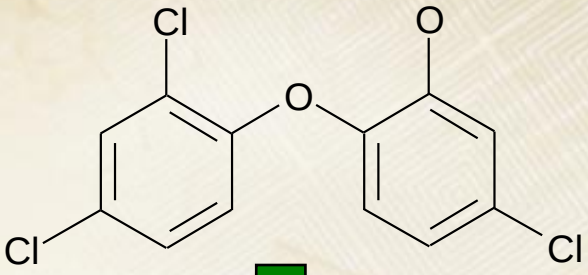
Data set – Toxicity Data

FDA classes	Definition	CAESAR Binary class	Total compounds
Category A	Negative human studies	Non developmental toxicant	91
Category B	Negative animal studies No human studies executed OR Positive animal studies Negative human studies		
Category C	Postive animal studies No human studies executed OR No studies at all	Developmental toxicant	201
Category D	Postive human studies		
Category X	Animal OR human studies show abnormalities AND/OR Evidence of foetal risk based on human experience		
			292

Molecular descriptors

DATA SET

2D structures



SOFWARE

- **MDL QSAR**
- **Dragon**
- **EPA (Free software)**
- **ACD/logD**
- **Pallas**
- **KowWIN**

2D descriptors families were computed and tested

Constitutional/information descriptors: molecular weight, number of chemical elements, number of H-bonds or double bonds, …

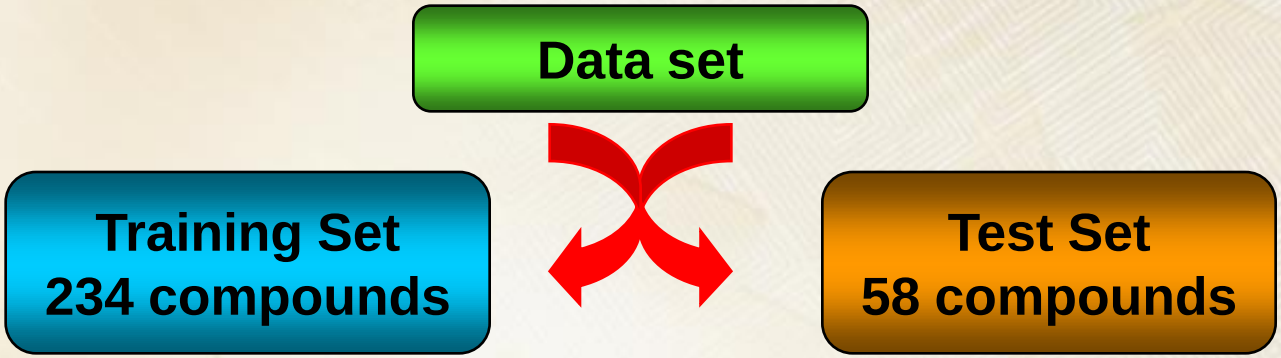
Physicochemical descriptors: lipophilicity, polarizability, …

Topological descriptors: atomic branching and ramification.

Training / Test sets selection

Set separation in rational and objective way based on chemical composition (atomic fragments)

Training set / test set ratio = 4 : 1



Building the prediction models

* **Enough compounds**
* Representative **molecular distribution**
* Representative **toxicity data**

Evaluating the prediction ability

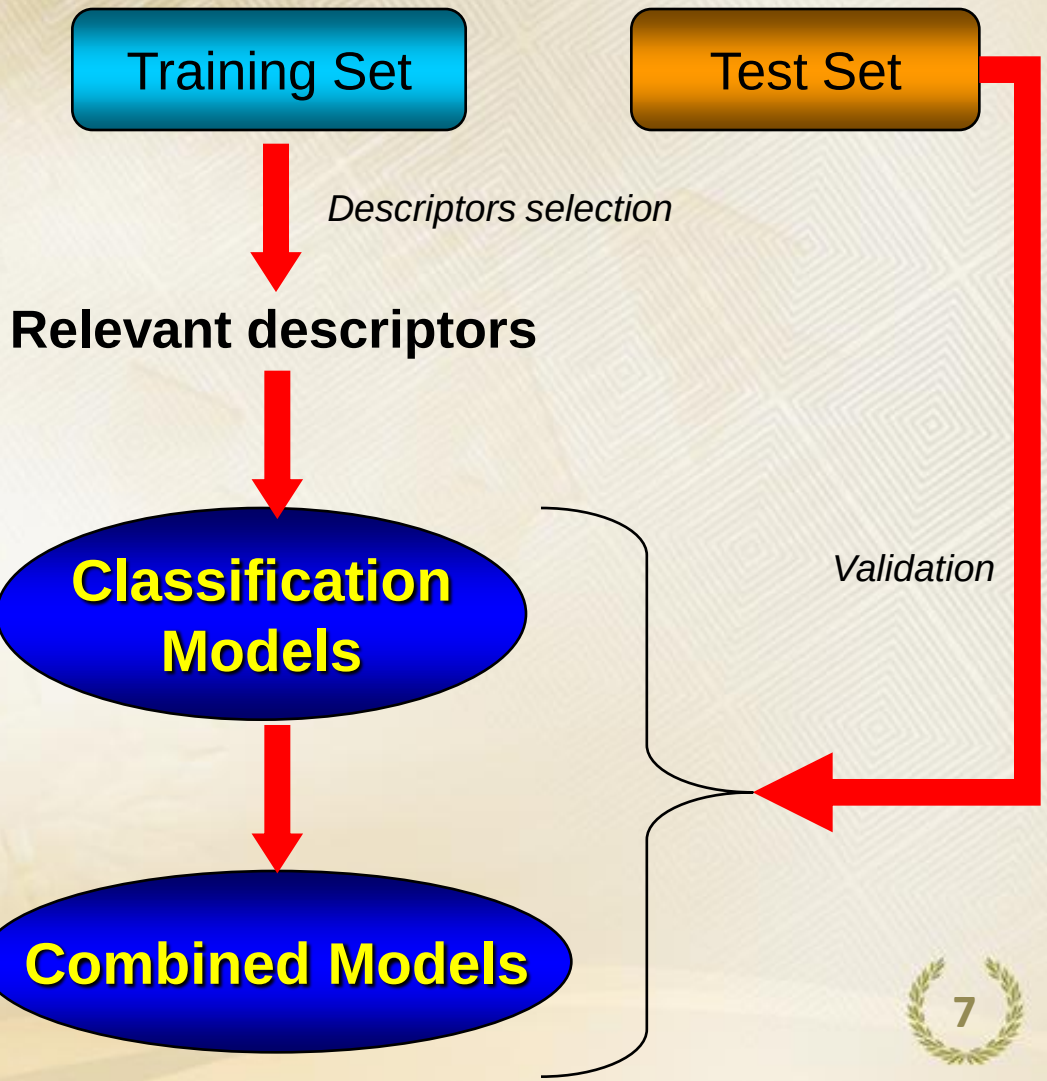
Compounds never used in the modelling process

Model development

METHODS

Descriptors Selection:
HSA
CfsSubsetEval

Model development:
AFP
GMDH
Tree Random Forest
MLP
Back propagation CO-NN



Validity and predictivity

- Battery of statistical checks, internal and external validation
- Attention to False Negatives (FN)
- Models optimized to reduce FN: REACH specific models
- Models using a low number of molecular descriptors

Results of DT modelling

Method	Nb des.	Des. Type	Training					Test		
			A	LOO	LSO	SE	SP	A	SE	SP
AFP	6	EPA	87		72	93	74	86	90	82
Tree Random Forest	8	EPA	100	74	76	100	100	81	88	65
Tree Random Forest	13	EPA	100	74	75	100	100	86	98	59
Tree Random Forest_S42	30	EPA	99	79	77	100	97	86	90	77
MLP+BP	8	MDL	85	76	77	90	73	83	88	71
GMDH NN	8	EPA	82	82		81	85	71	73	65
GMDH CO-NN	5	EPA	82	82		94	57	83	98	47
GMDH CO-NN (4 models)	13	EPA	87	87		96	68	86	100	53
GMDH NN (3 models)	16	EPA	86	86		86	86	79	88	59

Very good classification results for these models

A(Training)=82-100%; A(Test)= 71-86%

CV= about 75%

Model performance evaluation (1)

Validation statistics derived from the AFP model by using ONLY 6 EPA des.

MODEL 1

	TOTAL	Training	Test
Accuracy	87	87	86
Cross-validation (LSO)		72	
Nb unpredicted compounds	1	0	1
Total compounds	291	234	57
Accuracy	87	87	88
False Positive Rate	24	26	18
False Negative Rate	8	7	10
Postive Predictive Value	89	89	92
Negative Predictive Value	82	83	78
Sensitivity (class Developmental Toxicant)	93	93	90
Specificity (class Non toxicant)	76	74	82

Model performance evaluation (2)

Validation statistics derived from the DT_MN_EPA6 (MN) model by using 13 EPA descr.

Implemented MODEL 2

	TOTAL	Training	Test
Accuracy	97	100	86
Cross-validation (LSO)		75	
Nb unpredicted compounds	8	0	8
Total compounds	292	234	58
Accuracy	97	100	86
False Positive Rate	8	0	41
False Negative Rate	3	0	2
Postive Predictive Value	97	100	85
Negative Predictive Value	99	100	91
Sensitivity (class Developmental Toxicant)	99	100	98
Specificity (class Non toxicant)	92	100	59

Conclusion

- New integrated models for Developmental toxicity have been developed.
- All the models were statistically evaluated using strict criteria.
- Better performances than available models
- Focus on REACH:
  - Experimental data according to guidelines
  - Quality check (chemical structures)
  - Reproducibility
  - Transparency
  - False negatives minimized