WORKSHOP ON QSAR MODELS FOR REACH

Mario Negri Institute, Milan, Italy - March 10-11, 2009

SIXTH FRAMEWORK PROGRAMME

Emilio BENFENATI

Istituto di Ricerche Farmacologiche Mario Negri

Technical aspects behind the CAESAR models

http://www.caesar-project.eu/

DATA USED FOR CAESAR

- The preferred data were from *official protocols*
- ***Data of high quality***: if a model should be used for *regulatory purposes*, data have to be *carefully checked*

poor data -> poor models

- All models should ideally use *high quality data*
- Some **approaches** are more robust, others very dependent on data quality
- Check of data with *other data sources*
- Particular attention to *outliers*
- ***Double check*** all chemical structures
- CAESAR spent **one year just checking data**

QUALITY CONTROL

- **Some databases/sources are very good:**
  - **DSSTox (US EPA sources): *less than 1% errors***

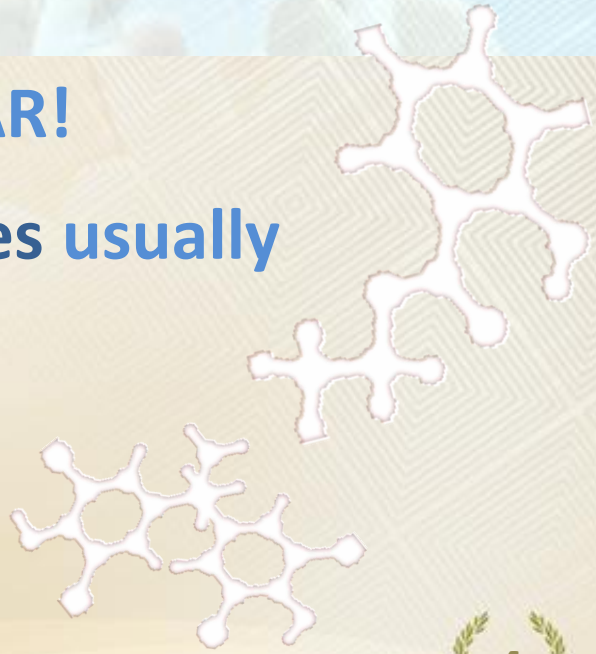
- **Still, we found errors there (*acknowledged*)**
  - ***Errors* may be:**
    - **wrong units (mg / µg; mg/mmol; </>)**
    - **wrong connectivity**
    - **wrong isomers**
    - **mixture**
    - **degradation/hydrolysis in the media**

Check of Chemical Structures

- Many errors, unsuitable data may be more than 20% of the structures in recent sources
- This affects not only QSAR, but all REACH

Always check, not only for QSAR!

Attention to chemical structures usually receives little attention



Chemical Information

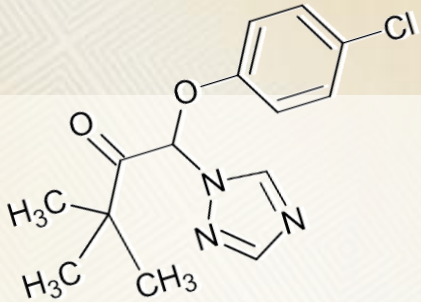
- After careful check of the chemical structure we extracted CHEMICAL INFORMATION

- Let's see the simple example of EPI Suite ™

$$Tox = 1.32 * LogP + 0.23$$

$$Tox1 = 0.55*des1 + 0.36*des2 + 0.29*des3 + 0.64*des4 - 0.47*des5 - 1.56*des6 - 0.53*des7 + 0.27*des8 + 0.55*des9 + 0.50*des10 + 0.23$$

KowWin (LogKow) Log P Calculation:



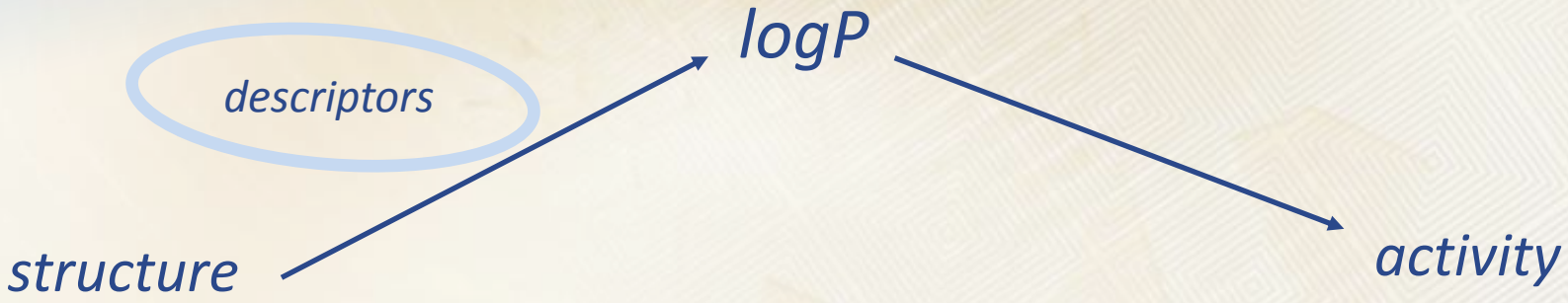
SMILES : CC(C)(C)C(=O)C(Oc1ccc(Cl)cc1)n2cncn2
CHEM : Triadimefon
MOL FOR: C14 H16 CL1 N3 O2
MOL WT : 293.76

TYPE	NUM	LOGKOW v1.66 FRAGMENT DESCRIPTION	COEFF	VALUE
Frag	3	-CH3 [aliphatic carbon]	0.5473	1.6419
Frag	1	-CH [aliphatic carbon]	0.3614	0.3614
Frag	8	Aromatic Carbon	0.2940	2.3520
Frag	1	-CL [chlorine, aromatic attach]	0.6445	0.6445
Frag	1	-O- [oxygen, one aromatic attach]	-0.4664	-0.4664
Frag	1	-C(=O)- [carbonyl, aliphatic attach]	-1.5586	-1.5586
Frag	3	Aromatic Nitrogen [5-member ring]	-0.5262	-1.5786
Frag	1	-tert Carbon [3 or more carbon attach]	0.2676	0.2676
Factor	1	-N-C-O- structure correction	0.5494	0.5494
Factor	1	-C-CO-C-O- structure correction	0.5000	0.5000
Const		Equation Constant		0.2290
			Log Kow =	2.9422

LogKow Estimated Log P: 2.94

EPI SUITE VS CAESAR

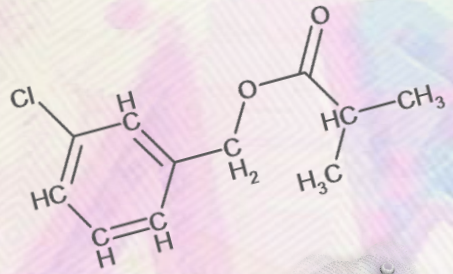
EPI Suite ™

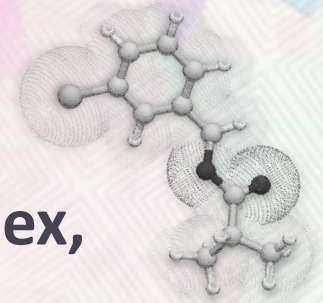


CAESAR

Chemical Information

- Chemical descriptors
- Fragments
- 2D preferred (3D tested, same results)
- 3D require manual optimization, more complex, longer, less suitable to automation

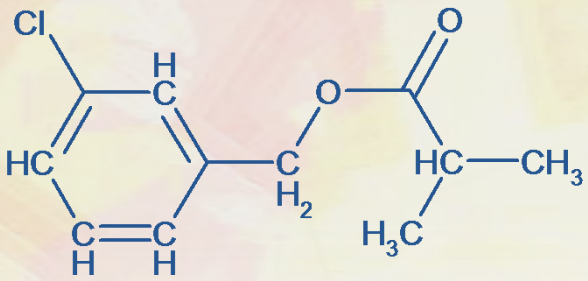




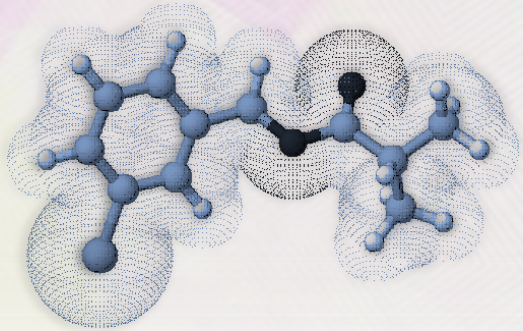
Reproducibility
is important for **REGULATORY PURPOSES**

PROCEDURE TO CALCULATE DESCRIPTORS

2D descriptors
(no optimization required)



3D descriptors
(optimization required)



Molecular Descriptors

- many descriptor families were tested:

Constitutional / information descriptors

molecular weight, number of chemical elements, number of H-bonds or double bonds, …

Physicochemical descriptors

lipophilicity, polarizability, …

Topological descriptors

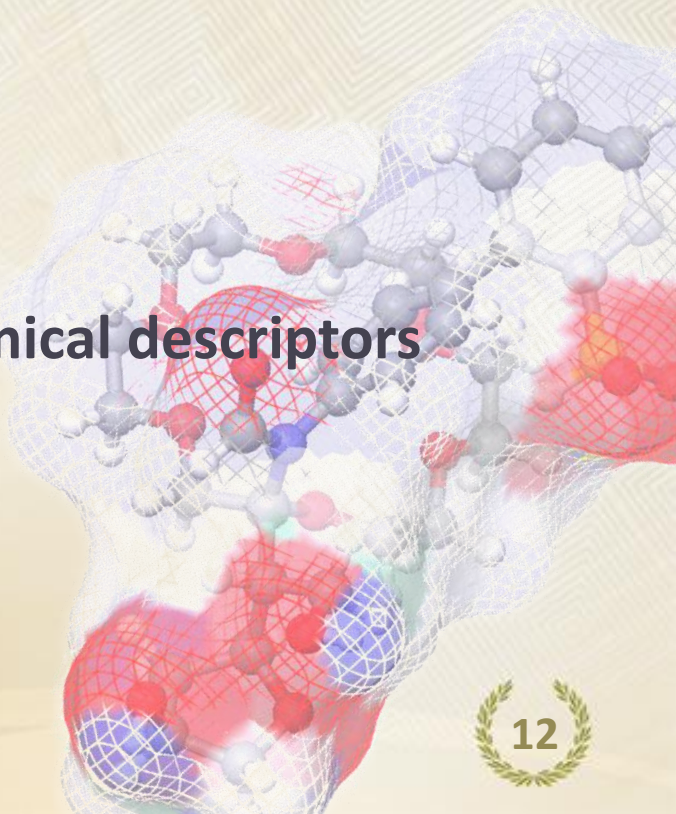
atomic branching and ramification

Electronic, geometrical and quantum-chemical descriptors

Fragmental / structural keys

defining Boolean (bitmap) arrays

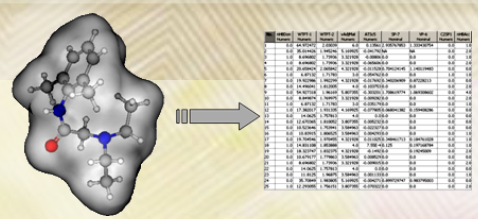
… … …



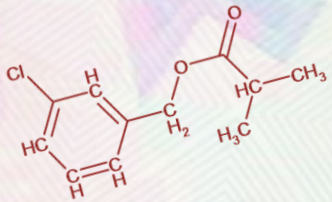
QSAR Selection Criteria

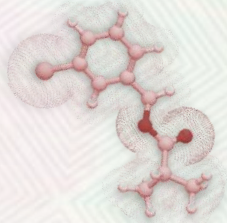
- *Inorganic compounds*, *complexes* and *mixtures* were discarded
- *Impure compounds* were discarded
- *Mixtures* were generally discarded
  - Exceptions: *stereoisomers*, *tautomers*
- *Organometallics* were retained
  - but often problematic (few examples)
- *Salts* and *hydrates* kept in neutral and anhydrous forms
  - see POSTERS
- *Chirality* was not taken into consideration

PROGRAMS TO CALCULATE DESCRIPTORS



- We used a series of programs, commercial and freely available
- Models were checked with all of them to identify the best
- Results vary, and specified for each model
- Use of the software and version specified





PUBLIC & COMMERCIAL TOOLS

- **when we started**
  most reliable *programs* for descriptors were *commercial*

- **In the CAESAR contract we specified this:**
  "use of commercial tools for descriptors"

- **Very recently we started a collaboration**
  with US EPA, for the use of their public program

ADVANTAGES OF COLLABORATION
CAESAR - US EPA

ECONOMIC
public program

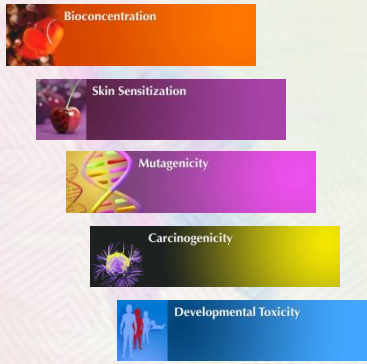
FULL CONTROL OF THE MODEL
no changes of versions

POSSIBLE IMPLEMENTATION INTO A UNIQUE TOOL
from structure to prediction

TRANSPARENCY

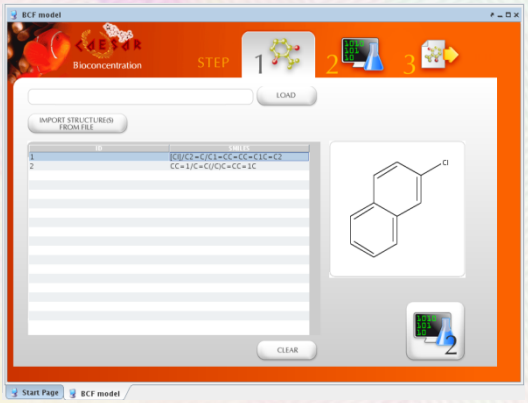
CAESAR MANDATE: ALGORITHM

- according to the CONTRACT
  - 5 MODELS
  - ALGORITHM, not descriptors



We did MORE

- \+ More than 5 models
- \+ Applets
- \+ Single tool in some cases



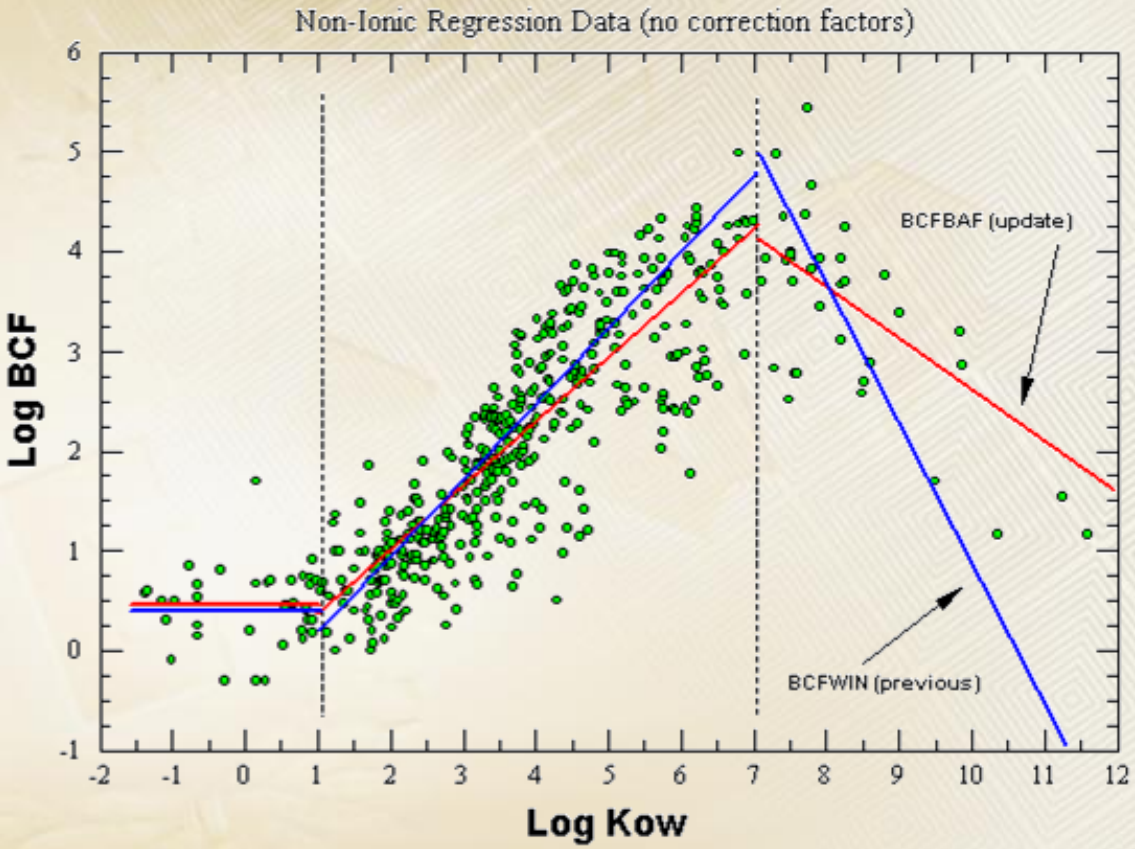
We are working to make all models as *SIMPLE APPLETS*

ALGORITHMS

- *Confusion* with tools to develop we model and the final algorithm
- Our *final models* are sequences of rules / simple equations
- *Advanced tools* used to select descriptors
- *Selection* can be done manually or with mathematical tools (PCA/GA)

EPI SUITE FOR BCF

- *Non linear, empirical, statistical*
- y = ax + b



LINEAR / NON-LINEAR MODELS

- *we used both linear and non-linear models to develop algorithms*

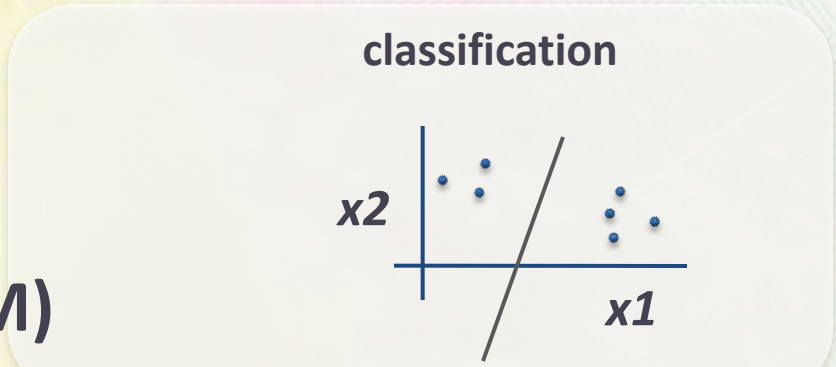
- **ALGORITHM**

  Classifiers

  Regressions

ALGORITHMS : CLASSIFIERS

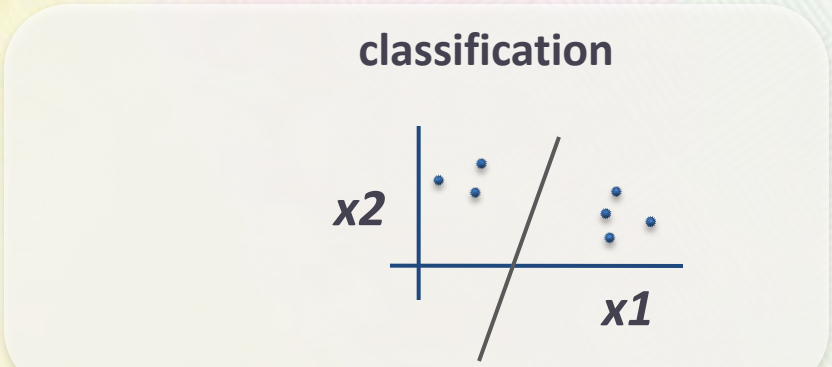
- Discriminant Analysis
- CART
- KNN
- Fuzzy logic
- Bayesian
- Self Organizing Map (SOM)
- Support Vector Machine (SVM)



ALGORITHMS : REGRESSIONS

- Multivariate Analysis (MVA)
- Partial Least Squares (PLS)
- Neural Networks (NN)
- Other algorithms

  (PCA, Genetic Algorithms)



EXPLICIT & IMPLICIT KNOWLEDGE

- We used **different tools**, combined where possible
- We used *explicit knowledge* (**rules from human experts**) combined with *implicit knowledge* (**statistical methods**)
- We had a *continuous flow of information and data* between partners: **cross-fertilization**
- *Different partners* worked on the *same endpoint* and **collaborated**

MULTIPLE QSAR APPROACHES

- **REACH promotes the use of all information sources**

- **We used a palette of many QSAR tools with different *data sets*, different *chemical information*, different *algorithms***
- **More than 5 models ready**
- **Many models will be exploited and implemented soon**
- **More studies on *outliers*, *reasons* and *mechanisms* will be added**

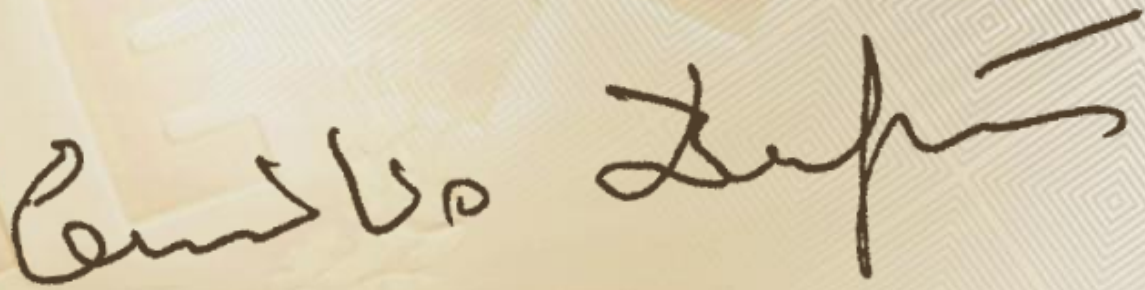
CAESAR

WORKSHOP ON QSAR MODELS FOR REACH

Mario Negri Institute, Milan, Italy - March 10-11, 2009

SIXTH FRAMEWORK PROGRAMME

GRAZIE!



MARIO NEGRI
ISTITUTO DI RICERCHE FARMACOLOGICHE