Applicazione delle metodologie QSAR a problematiche ambientali di inquinanti organici

Prof. Paola Gramatica

QSAR and Environmental Chemistry Research Unit
DBSF - University of Insubria - Varese
http://www.qsar.it
QSAR

Quantitative Structure - Activity Relationships
Quantitative Structure-Activity Relationships (QSAR) and Quantitative Structure-Property Relationships (QSPR)

“The structure of a chemical influences its properties and biological activity”

“Similar compounds behave similarly”

(Hansch 1964)

Activity or Property = f (Structure)

It is possible to find a relationship (f) between Structure and behavior (Activity or Property) of a chemical

PREDICTED DATA
Chemicals

Natural products

Xenobiotics

Physico-chemical properties

Environmental fate and behavior
- degradation
- persistence
- bioaccumulation
- partitioning

Biological Activity
- toxicity
- mutagenicity
- carcinogenicity
- endocrine disrupt
THE CHEMICAL UNIVERSE

22.000.000 in C.A.S.
100.000 on market

EINECS TSCA

5% known data

NEW 1.000.000 / year

NEW 2.000 / year

Environmental fate?

Human effects?

experiments

Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)
(Q)SAR History

Alkane m.p. and b.p. (Cros, 1863)

Alcohol water solubility

PHYSICO-CHEMICAL PROPERTIES

BIOLOGICAL ACTIVITY

STRUCTURE/PROPERTIES

Alcohol toxicity (Meyer-Overton 1899-1901)

part. coeff. oil/water

Log P

PHYSICO-CHEMICAL PROPERTIES

BIOLOGICAL ACTIVITY

STRUCTURE/PROPERTIES

(Hansch 1964)
Classical Hansch equation:

\[ \text{“Toxicity”} = a + b \log P + c E + d S \]

- **logP** or **log Kow**, partition coefficient between octanol and water: 
  - hydrophobicity term
- **E** electronic term
- **S** steric term related to bulk and shape

The probability or ability of the chemical to reach the target site

The possibility of the chemical to interact with the target and to be active

**Congenericity principle**
M: experimental measures of properties
A: experimental measures of activities
D: theoretical procedures for descriptors
R1, R2, R3: mathematical relationships

Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)
THE 3 NECESSITIES:

GOOD INPUT DATA

High-quality experimental data as input data to find the Structure-Activity Relation

MEANINGFUL STRUCTURAL INFORMATION

Good representation of the chemical structure: molecular descriptors

PREDICTIVE MODELS

Quantitative models with validated predictive performances
Experimental data set

There is a need for a “limited” number of HIGH-QUALITY experimental data on which to develop QSAR models!

The models will only be as good as the data used to develop them!

“Garbage in, garbage out”

NEEDS FOR EXPERIMENTAL DATA:

- AS NUMEROUS AS POSSIBLE
- CORRECT
- REPRESENTATIVE
- HOMOGENEOUS

(ideally, same lab, same method)
M: experimental measures of properties

A: experimental measures of activities

D: theoretical procedures for descriptors

R1, R2, R3: mathematical relationships
The “magic” molecular descriptor

Log P (or Kow)

“35 years of (ab)using of log P for everything modelling is enough! “

(R. Schwarzenback)

SETAC 2000

FROM PARTITION PROPERTY TO ACTUAL MOLECULAR STRUCTURE...
MOLECULAR DESCRIPTORS

Representation of a chemical by numerical indices

Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)
M: experimental measures of properties
A: experimental measures of activities
D: theoretical procedures for descriptors
R1, R2, R3: mathematical relationships

Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)
EXPLORATIVE ANALYSIS:
- Principal Component Analysis
- Cluster Analysis

CLASSIFICATION METHODS:
- Classification Tree (CART)
- Discriminant Analysis
- Neural Networks

REGRESSION METHODS:
- Multivariate Linear Regression (MLR)
- Partial Least Squares Regression (PLS)

Quantitative models for qualitative responses

Quantitative models for quantitative responses

Chemometric Methods

Chemometric Methods

Molecular Descriptors

Experimental Data

Chemicals

$\begin{array}{cccc}
 x_1 & x_2 & \ldots & x_n \\
 \hline
 \end{array}$

Training Set

$Y$
DATA SET

SPLITTING

- Dimension
- Chemical composition

TRAINING SET

FITTING $R^2$

TEST SET

EXTERNAL VALIDATION $Q^2_{\text{EXT}}$

REGRESSION MODEL

INTERNAL VALIDATION $Q^2_{\text{LOO}}$
$Q^2_{\text{LMO}}$

PREDICTABILITY

NEW DATA
MOLECULAR DESCRIPTORS

SELECTION

Y = f (selected descriptors)

QSAR MODEL

Exhaustive description

Model with relevant information

Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)
LIMITATIONS OF QSAR MODELS

- Statistical quality
  - Fitting $R^2$
  - Predictability $Q^2$
- Chemical domain
- Outliers
- Prediction reliability

Exp. response vs Pred. response
CHEMICALS — EXPERIMENTAL DATA — MOLECULAR DESCRIPTORS

FITTING

Y

QSAR QSAR

MAXIMUM PREDICTIVE POWER

MODEL

REVERSIBLE DECODING

NEW CHEMICALS — ??? PREDICTION — MOLECULAR DESCRIPTORS

Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)
APPLICATIONS of QSAR PREDICTIONS

- Filling of data gaps
- Validation of experimental data
- Screening, ranking and priority setting
- Highlighting chemicals of concern (also before their synthesis)

PRIORITIZATION OF RISK ASSESSMENT

Optimize industry resource allocation

Minimize animal testing
## Environmental Parameters

### Priority setting / Risk Assessment

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Quality of QSAR models</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physico-chemical data</strong></td>
<td></td>
</tr>
<tr>
<td>m.p.; b.p.; vapour pressure; Henry law constant; water solubility; partition coefficients (Kow, Koc, …).</td>
<td><strong>OPTIMUM</strong></td>
</tr>
<tr>
<td><strong>Environmental fate and pathways</strong></td>
<td></td>
</tr>
<tr>
<td>chemical-, photo- and bio-degradation; bioaccumulation; compartment partitioning.</td>
<td><strong>HIGH</strong></td>
</tr>
<tr>
<td><strong>Ecotoxicity</strong></td>
<td></td>
</tr>
<tr>
<td>algae; <em>Daphnia</em>; fish; …</td>
<td><strong>MEDIUM-HIGH</strong></td>
</tr>
<tr>
<td><strong>Mammalian toxicity</strong></td>
<td></td>
</tr>
<tr>
<td>skin-, eyes-, oral-, inhalation acute toxicity; mutagenicity; carcinogenicity; toxicity to reproduction system;...</td>
<td><strong>MEDIUM</strong></td>
</tr>
</tbody>
</table>
### QSAR in U.S.

**Since 1979/80 wide use and “abuse”**

<table>
<thead>
<tr>
<th><strong>EPA / OPPT</strong></th>
<th><strong>Office of Pollution Prevention and Toxics</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TSCA</strong></td>
<td>Toxic Substances Control Act inventory (~75,000 chem.)</td>
</tr>
<tr>
<td><strong>NCP</strong></td>
<td>New Chemicals Program (PMN with QSAR data)</td>
</tr>
</tbody>
</table>

### QSAR in E.U.

**Since 1992/93 but, so far, limited use**

- **EC Regulation on Evaluation and Control of Risks of Existing Substances**

<table>
<thead>
<tr>
<th><strong>IPS</strong></th>
<th>Informal Priority Setting method</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EURAM</strong></td>
<td>Europe Union RAning Method</td>
</tr>
</tbody>
</table>
WHITE PAPER on the Strategy for a Future Chemicals Policy (Feb 2001)

• Same regulation for new and existing chemicals (1:15)

• Responsibility from authorities to industries for testing and risk assessment

• REACH system: Registration Evaluation Authorisation of Chemicals

- Registration by companies for > 1 t prod (30,000) into 2005 1000 t (HPV), into 2012 all

- Evaluation of information by authorities for > 100 t (5000) into 2008

- Authorisation for carcinogenic, mutagenic, toxic to reproduction and POPs
QSAR in WHITE PAPER

Art. 3.2 ….”to keep animal testing to a minimum”

- Development and validation of alternative methods
ECVAM (Europ. Centre for Validation of Alternative Methods) – JRC Ispra

- Inclusion in the Community legislation and OECD Test Guidelines Programme for international recognition

Task Force of Experts in QSAR

- Particular research efforts for developing and validating modelling (e.g. QSAR) and screening methods for assessing the potential adverse effects of chemicals.

Setubal Principles
Classi di composti studiate

- **POP** (*Persistent Organic Pollutants*), **PBT** (*Persistent Bioaccumulative Toxics*)
- **VOC** (*Volatile Organic Compounds*) e **HPV** (*High Production Volume*)
- Pesticidi: insetticidi, erbicidi, ...
- Idrocarburi aromatici policondensati (PAH)
- Bifenili policlorurati (PCB), diossine
- Benzeni e fenoli sostituiti
- Prodotti Industria Chimica Italiana (per FEDERCHIMICA)
- Lista di Priorità 1 della EU
Applicazioni in campo ambientale

Predizione di proprietà chimico-fisiche per studi di ripartizione nei comparti ambientali:
- bioconcentrazione (BCF)
- volatilità (log Koa, log H, Vp)
- coefficiente di adsorbimento nel suolo (log Koc)
- indici di mobilità (leaching…….)

Predizione di parametri di persistenza ambientale:
- reattività atmosferica (costanti di velocità di reazione con radicali OH, NO$_3$ ed O$_3$)
- indici di persistenza ambientale (emivite)
- biodegradabilità ……..

Predizione di attività biologiche:
- tossicità
- mutagenicità ……..
Analisi delle Componenti Principali (PCA)

La PCA è una analisi esplorativa di dati multivariati.

Le Componenti principali sono:
• Combinazioni lineari dei dati originali
• Ordinate secondo le direzioni di massima varianza (PC1, PC2..)
• Non correlate
• Sono quindi nuove variabili con le quali si condensa e “pulisce” l’informazione contenuta nei dati originali
• Rappresentano macroproprietà dell’insieme dei dati originali

La PCA consiste in una rotazione nello spazio dei dati originali in modo che le singole componenti siano tra loro ortogonali

I dati vengono così “visti” in un diverso sistema di riferimento Secondo visuali controllate per qualità e quantità dell’informazione rappresentata
Reaction rate constants for the degradation by Tropospheric Oxidants: $\text{OH}^\cdot, \text{NO}_3^\cdot$ radicals and Ozone
GLOBAL ATMOSPHERIC PERSISTENCE INDEX (ATPIN) 65 VOCs

PC1 score = ATPIN
(Atmospheric Persistence Index)
65 objects 4 variables
R²=93.76% Q²loo=92.53
Q²LMO(50%)=91.44% RMS=0.163

DESCRIPTORS : (in order of significance)
• HOMO: highest occupied molecular orbital (nucleophilicity)
• nBnz: number of aromatic rings
• AMW: average molecular weight
• DELS: molecular electropological variation (charge distribution)

Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)
Example: QSAR Models for Degradation by NO$_3^-$ (114 chemicals)

<table>
<thead>
<tr>
<th>Obj.Tr.</th>
<th>Obj.Test</th>
<th>Var.N.</th>
<th>VARIABLES</th>
<th>$R^2$</th>
<th>$Q^2_{LOO}$</th>
<th>$Q^2_{LMO(50%)}$</th>
<th>$Q^2_{ext}$</th>
<th>RMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>114</td>
<td>3</td>
<td>3</td>
<td>HOMO nBnz MATS1m</td>
<td>92.9</td>
<td>92.3</td>
<td>92.1</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>37</td>
<td>3</td>
<td>HOMO nBnz MATS1m</td>
<td>90.3</td>
<td>91.2</td>
<td>89.6</td>
<td>95.9</td>
<td></td>
</tr>
</tbody>
</table>

DESCRIPTORS (in order of significance):

- HOMO: highest occupied molecular orbital (nucleophilicity of molecules)
- nBnz: number of aromatic rings
- MATS1m: 2D-autocorrelation of Moran (atomic distribution)


Paola Gramatica - QSAR and Environmental Chemistry Research Unit
DBSF - INSUBRIA University - (Varese - ITALY)
Principal Component Analysis
Cum. E.V.% = 95.3% (PC1 = 80.9%)

PC1 = ATPIN

-\log k(O_3)
-\log k(NO_3)
-\log k(OH)

Atmospheric Persistence

Exp. + Pred. (399 obj.)
Exp. (65 obj.)


Paola Gramatica - QSAR and Environmental Chemistry Research Unit
DBSF - INSUBRIA University - (Varese - ITALY)
QSAR Modelling of “enlarged” GLOBAL ATMOSPHERIC PERSISTENCE INDEX

PC1 score = \textbf{ATPIN} \ (399 chemicals experimental + predicted data)

<table>
<thead>
<tr>
<th>Obj.Tr.</th>
<th>Obj.Test</th>
<th>Var.N.</th>
<th>\textbf{VARIABILI}</th>
<th>(R^2)</th>
<th>(Q_{\text{LOO}}^2)</th>
<th>(Q_{\text{LMO}(50%)}^2)</th>
<th>(Q_{\text{ext}}^2)</th>
<th>RMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>399</td>
<td>3</td>
<td>HOMO nBnz BEHe4</td>
<td>93.3</td>
<td>93.2</td>
<td>93.2</td>
<td>0.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>255</td>
<td>174</td>
<td>HOMO nBnz BEHe4</td>
<td>93.7</td>
<td>93.5</td>
<td>93.4</td>
<td>92.7</td>
<td>0.41</td>
<td></td>
</tr>
</tbody>
</table>

\textbf{DESCRIPTORS (in order of significance):}

- \textbf{HOMO}: highest occupied molecular orbital (nucleophilicity of molecules)
- \textbf{nBnz}: number of aromatic rings
- \textbf{BEHe4}: weighted by electronegativity (charge distribution)

Paola Gramatica - QSAR and Environmental Chemistry Research Unit
DBSF - INSUBRIA University - (Varese - ITALY)
Ranking of pesticides for environmental distribution, based on PCA

Principal Component Analysis (PCA) ON CHEMICAL-PHYSICAL PROPERTIES OF 54 PESTICIDES

Cum. E.V. = 94.6% (PC1 = 70.1%)
Clustering of pesticides for environmental distribution in 4 \textit{a priori} classes

![Dendrogram](image)

- **1: Not-volatile/medium comp.**
- **2: Volatile comp.**
- **3: Soluble comp.**
- **4: Sorbed comp.**
Definition of 4 a priori classes of pesticides for environmental distribution

PCA ON CHEMICAL-PHYSICAL PROPERTIES OF 54 PESTICIDES
Cum. E.V. = 94.6% (PC1 = 70.1%)

CLA 1 = NOT-volatile COMP.
CLA 2 = VOLATILE COMP.
CLA 3 = SOLUBLE COMP.
CLA 4 = SORBED COMP.
Classification of 54 pesticides for environmental distribution

Classification Tree

DESCRIPTORS:
- MW: molecular weight (size)
- nHD: number of donor atoms in Hydrogen bonds
- J: Balaban topological index

Environmental Partitioning for 135 pesticides

The “FOUR-LEAVES” approach

(P.Gramatica and A.Di Guardo, Chemosphere, 2002, 47, 9547-956)

PC1 score = LIN Leaching Index  PC2 score = VIN Volatility Index

Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)
Environmental Partitioning Indexes Regression

**LIN (Leaching Index)** =

\[-3.04 - 0.96 \, nX - 2.28 \, nNO + 3.42 \, Ms\]
\[-1.74 \, ICR - 0.45 \, nS\]

n = 135  \( R^2 = 87.0\% \)  \( Q^2_{LOO} = 85.8\% \)
\( Q^2_{LMO} = 85.7\% \)  SDEP = 0.68  SDEC = 0.65

**VIN (Volatility Index)** =

\[2.35 - 1.58 \, HY - 0.23 \, X_v - 0.28 \, nBM - 0.40 \, nCIC - 1.42 \, ASP + 0.97 \, IVDE\]

n = 135  \( R^2 = 77.1\% \)  \( Q^2_{LOO} = 74.1\% \)
\( Q^2_{LMO} = 72.6\% \)  SDEP = 0.66  SDEC = 0.63


Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)
**WQO - Water Quality Objectives**

**EEC Priority List 1**

toxicity test on algae, Daphnia, fish

---

**TOXICITY IN DAPHNIA**

\[
\text{Log } \frac{1}{\text{EC50}} = -3.57 + 4.05 \text{nP} - 0.39 \text{nHA} + 1.02 \text{IDM} + 0.67 \text{E1m}
\]

\( n = 94 \quad R^2 = 84.2\% \quad Q^2_{\text{LOO}} = 82.1\% \quad Q^2_{\text{LMO}} = 81.7\% \)

nP: n. of phosphorous atoms  

nHA: n. of H bond acceptors  

IDM: mean inf. cont. on the dist. magn.  

E1m: distribuz. atomica  

dir-WHIM descriptor

---

**PRINCIPAL COMPONENT ANALYSIS (PCA)**

- all toxicity data available for 37 chemicals (E.V.: 90%)  
- experimental + predicted data for 97 chemicals (E.V.: 93.7%)

*Ecotox and Environ Safety, 49, (2001) 206-220*

---

**CLASSIFICAZIONE CART**

Obj. n.: 125  
Selected var.: ZM1 - CHI0

NoModel ER: 40.8\%  
ER: 7.2\%  
\( cv\text{ER}: 15.2\% \)

ZM1: Zagreb index  
CHI0: connectivity index

---

Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)
SCREENING of POPs for overall persistence based on half-life in air, water, soil

**HALF-LIFE in SOIL**

Log h.l. soil = -3.46 + 0.58 IDM + 0.99 E2m + 0.48 G2e

obj. = 30  $R^2 = 83.2\%$  $Q^2_{LOO} = 77.8\%$  $Q^2_{LMO} = 76.9\%$

IDM : mean inf. index on distance magnitude
E2m- G2e : directional WHIMs

**PRINCIPAL COMPONENT ANALYSIS (PCA)**

all half-life data available for 29 chemicals  (Cum. E.V.: 87.6\%)
experimental + predicted data for 91 chemicals  (Cum. E.V.: 79.5\%)

**OVERALL PERSISTENCE INDEX (PC1)**

PC1 = 9.22 + 3.14 AAC - 6.32 E2s - 17.49 E1e - 0.16 Tm

obj. = 91  $R^2 = 85.1\%$  $Q^2_{LOO} = 82.6\%$  $Q^2_{LMO} = 82.2\%$
Per avere minimo potenziale di LRT:

- Minimizzare la persistenza (PC1 in PCA delle emivite, fig. precedente)
- Minimizzare la mobilità (PC1 in PCA delle proprietà chimico-fisiche, es. a lato)

**MCDM: Multicriteria Decision Making**

\[
R(x) = \sum_{i=1}^{k} \epsilon_i f(x)
\]

**Funzione di utilità**

**Regressione e Classificazione**

CART (nC, E1u):

- \( MR_{cv} = 6.2\% \)
- No Model 42%

**Screening** dei POP per il Long Range Transport (LRT)

Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)
Acknowledgements

Dr. Ester Papa

Dr. Pamela Pilutti

Dr. Francesca Battaini

Dr. Fulvio Villa

http://dipbsf.uninsubria.it/qsar

http://www.qsar.it