An introduction to QSAR

Dr. Ioannis Nicolis
Faculté de Pharmacie
Université Paris Descartes

The central idea of QSAR

QSAR: Quantitative Structure Activity Relationships

A special case of

SPC: Structure Property Correlations

The biological activity of a molecule is a function of its physicochemical properties and of its structure

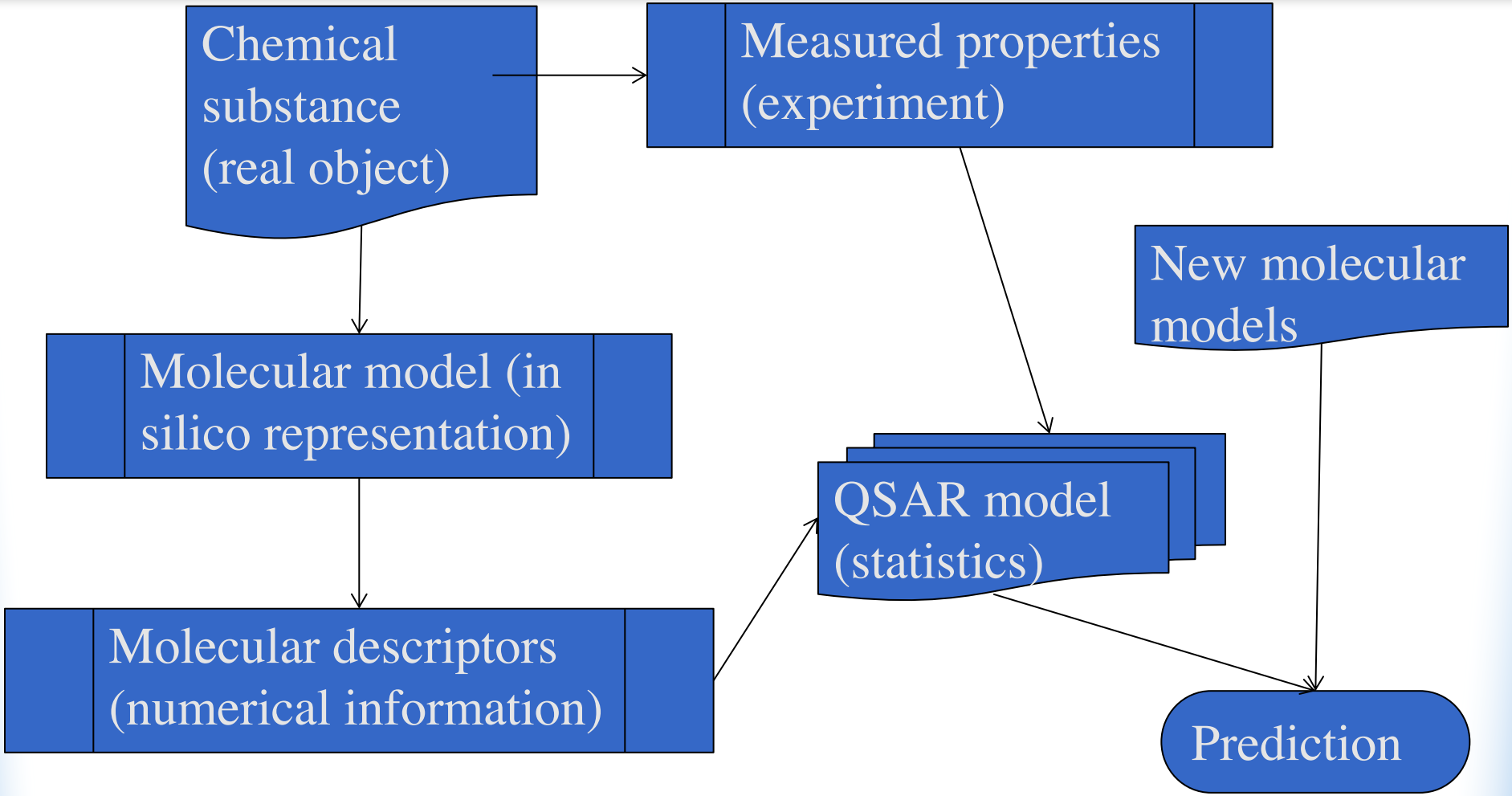
QSAR: what purpose?

- Understand which particular characteristics of a molecule have an influence on the measured activity
- Orient the conception of new compounds to optimize the property of interest
- Predict *in silico* the activity of new compounds

QSAR: what do you need?

- A series of molecules (usually of the same family) on which the property of interest has been measured
- Cheminformatics software to compute a "mathematical description" of the molecules (more on this later)
- Statistical software to build, refine and validate the predictive model

The procedure



Some history

- The precursors : “Physiological action” of a substance depends on its chemical structure: $\Phi = f(C)$
  - Crum-Brown and Fraser, 1868
- First applications :
  - Richet (1893): cytotoxicity depending on solubility
  - Meyer (1899), Overton, 1901: narcotic action depending on oil/water partition coefficients
  - Ferguson, 1939, Bell et Roblin (1942), Albert (1945) : effect of thermodynamic parameters and acid-base ionisation
- The substituent and steric effect, the σ ρ approach
  - Hammett (1935), Taft (1952)
- The founding of “Modern” QSAR
  - Hansch, Muir and Fujita (1962 et 1964)
  - Free and Wilson (1964)

QSPR/QSAR

- QSPR : Quantitative Structure–Property Relationships
  - The most known application: QSAR
    - QSAR : Quantitative Structure Activity Relationships
      - Activity: C, ID50, ED50, IC50, K_i, K_a, …
  - But you can encounter also the terms:
    - QSRR : Quantitative Structure–Retention Relationships
    - (QSRR : Quantitative Structure–Reactivity Relationships)
    - QSERR : Quantitative Structure–Enantioselective Retention Relationships
    - QSMR : Quantitative Structure–Metabolism Relationships
    - QSBR : Quantitative Structure–Biodegradation Relationships
    - QSER : Quantitative Structure–Extraction Relationships
    - …

The principle

Activity = f(structure)

- Two distinct problems
  - How to express in mathematical terms a chemical formula or a molecular structure
    - A computational chemistry and molecular modelling problem, using mathematics, topology and theoretical chemistry
  - What is the form of the link function and how to determine its parameters?
    - A statistics problem, needing special algorithms as most of the time number of variables largely outnumbers number of molecules

Introducing a molecule into an equation

A chemical formula or a molecular structure need to be "translated" in a numeric form in order to use them in mathematical equations.

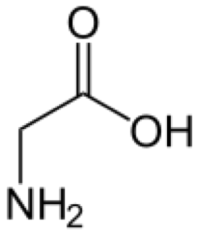
They can be expressed in terms of:

- Physicochemical properties
  - Solubility, lipophilicity, ionisation constants,…
  - Experimental or computed
- "Molecular descriptors"
  - Numerical parameters computed from the geometry, charge, functional groups… of the molecule
    - Bond topology, 3D geometry, chemical composition, partial charge distribution, aromaticity…

How to input a chemical formula in a computer: the SMILES notation

Simplified Molecular Input Line Entry Specification: SMILES

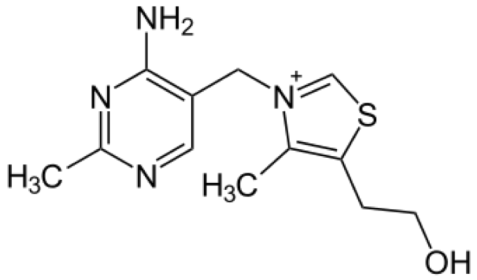
- A linear coding using letters. Hydrogens are omitted
  - Ethanol is coded CCO
- Double bonds are coded = and triple #
- Parentheses indicate branching
  - Glycine is coded C(C(=O)O)N



SMILES notation (continued)

- Cycles are coded using numbers to indicate that the atoms preceding the number are bonded
  - Cyclohexane is coded C1CCCCC1
- Lower case letters denote aromatic atoms
  - Benzene is coded c1ccccc1
- Charges are noted between square brackets
  - Vitamine B1 is coded:

    OCCc1c(C)[n+](=cs1)Cc2cnc(C)nc(N)2



- Stereochemistry is indicated with the symbols / \ @

SMILES in practice

- Obviously the conversion to and from SMILES is not performed by hand but using software
  - Local software
  - On line interfaces
- There are many equivalent SMILES notations for a molecule
- Two extensions
  - SMARTS: uses “wildcard” characters to code not a specific molecule but a family using patterns
  - SMIRKS: adds extra symbols to code chemical reactions
- SMILES is not the only way to code a molecule
  - International Chemical Identifier (InChI)

Molecular descriptors

Classification according to descriptor type

- Topological
  - Connectivity, molecular graphs, indices…
- Quantic
  - Energies HOMO and LUMO…
- Electrostatic
  - Total and partial charges, dipole moments…
- Geométric
  - Volume, solvent accessible surface…
- Physicochemical
  - Solubility, lipophilicity, pK_a…

Molecular descriptors

Classification according to molecule representation

- 0D
  - Depending only on counts derived from chemical formula
- 1D
  - Depending on one dimension representation
- 2D
  - Depending on atom connectivity
- 3D
  - Depending on 3D structure
    - i3D : structure of isolated molecule
    - x3D : structure of the molecule in a specific configuration (aligned in an active site for instance)

Examples

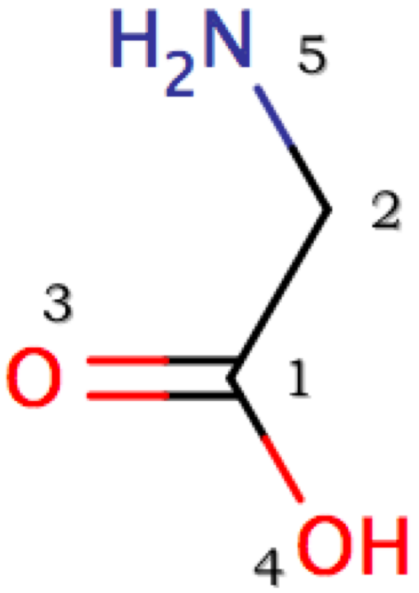
- 0D Descriptors
  - Molecular weight, number of atoms of a type, mean volume per atom…
- 1D Descriptors
  - Number of cycles, number of H-bond donors and acceptors…
- 2D Descriptors
  - Topological indices
- 3D Descriptors
  - Solvent accessible surface, gyration radius…

Topological indices

- A molecule can be described as a molecular graph
  - A mathematical object G=(V,E)
    - V: vertices (atoms), also called nodes
    - E: edges (bonds)
- From the molecular graph we compute two square matrices of dimension n x n (for a n atom molecule)
  - Connectivity matrix a_{ij} = 1 if atoms i and j are bonded, 0 otherwise
  - Distance matrix d_{ij} = distance (expressed as number of bonds) separating atoms i and j
- Other matrices can also be computed (degrees, edges,…)

Example: connectivity matrix of glycine

	1	2	3	4	5	δ_i
1	0	1	1	1	0	3
2	1	0	0	0	1	2
3	1	0	0	0	0	1
4	1	0	0	0	0	1
5	0	1	0	0	0	1

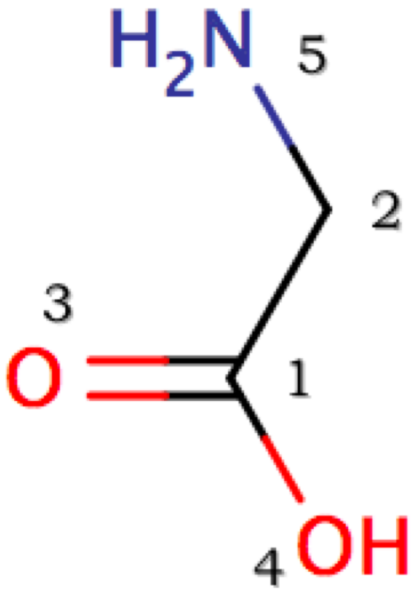


δ_i is the edge sum by line = number of bonds of atom *i*

Higher for atoms with many non-H substituents

Exemple: distance matrix of glycine

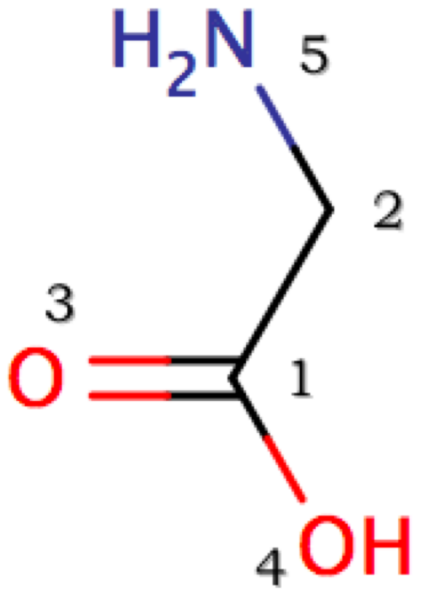
	1	2	3	4	5	s_i
1	0	1	1	1	2	5
2	1	0	2	2	1	6
3	1	2	0	2	3	8
4	1	2	2	0	3	8
5	2	1	3	3	0	9



s_i is the distance sum per line = higher for ”end” atoms, lower for ”middle” atoms

Example: degrees of glycine nodes

i	d_i
1	3
2	3
3	1
4	1
5	1



d_i is the degree of a vertex = number of non-hydrogen substituents

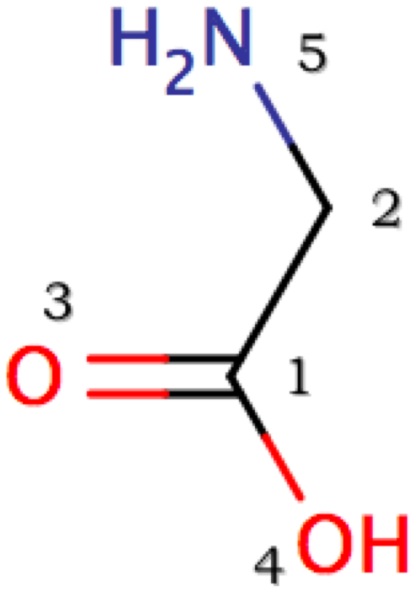
From molecular graphs to topological indices

- A large variety of topological indices can be computed from these graphs
  - Randic index
    - Harmonic sum of the geometric means of the node degrees for each edge
  - Wiener index
    - Average topological atom distance
  - Platt index
    - Sum of the edge degrees of a molecular graph
  - …

Randic index

- Randic index, computed from the vertices degree matrix

i—*j*	d_i	d_j
5—2	1	2
2—1	2	3
1–3	3	1
1—4	3	1



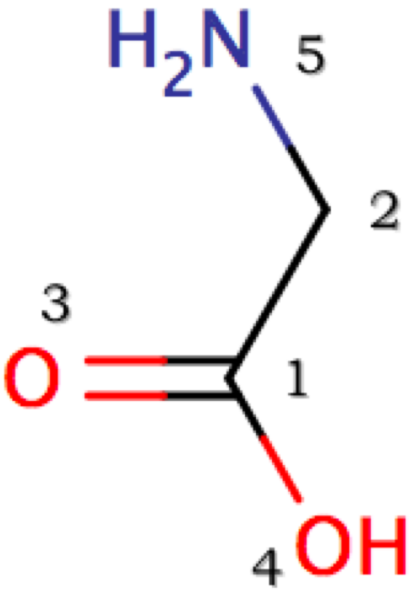
$$R(G) = \sum_E \frac{1}{\sqrt{d_i d_j}} = \frac{1}{\sqrt{1\times 2}} + \frac{1}{\sqrt{2\times 3}} + \frac{1}{\sqrt{3\times 1}} + \frac{1}{\sqrt{3\times 1}} \approx 2.27$$

Wiener index

- Wiener index, computed from the distance matrix

$$W(G) = \frac{1}{2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ i\neq j}}^{n} d_{ij} = \frac{1}{2}(5+6+8+8+9) = 16$$

	1	2	3	4	5	s_i
1	0	1	1	1	2	5
2	1	0	2	2	1	6
3	1	2	0	2	3	8
4	1	2	2	0	3	8
5	2	1	3	3	0	9



Kier & Hall molecular connectivity index (χ)

- χ index, takes into account σ electrons
- For each atom we compute

$\delta_i = \sigma - h$ where σ is the number of electrons in σ bonds and h the number of H bonded to atom *i*

- Then for the 0th order we compute:

$$c = \delta^{-\frac{1}{2}} \quad and\ the\ index \quad {}^0\chi = \sum_i c_i$$

Kier & Hall molecular connectivity index (χ)

- For the 1st order index we use the edges (bonds) instead of the nodes (atoms):

$$c_{ij} = \left(\delta_i \times \delta_j\right)^{-\frac{1}{2}} \quad and\ the\ index \quad {}^1\chi = \sum_{ij} c_{ij}$$

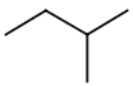
- We continue with higher orders taking into account pairs (or triplets…) of edges

Indexes in a series : example

	1	2	3	4	5	s_i
1	0	1	2	3	4	10
2	1	0	1	2	3	7
3	2	1	0	1	2	6
4	3	2	1	0	1	7
5	4	3	2	1	0	10

$$W = \frac{1}{2}(10 + 7 + 6 + 7 + 10) = 20$$

$$R = 2.41$$



	1	2	3	4	5	s_i
1	0	1	2	3	3	9
2	1	0	1	2	2	6
3	2	1	0	1	1	5
4	3	2	1	0	2	8
5	3	2	1	2	0	8

$$W = \frac{1}{2}(9 + 6 + 5 + 8 + 8) = 18$$

$$R = 2.27$$

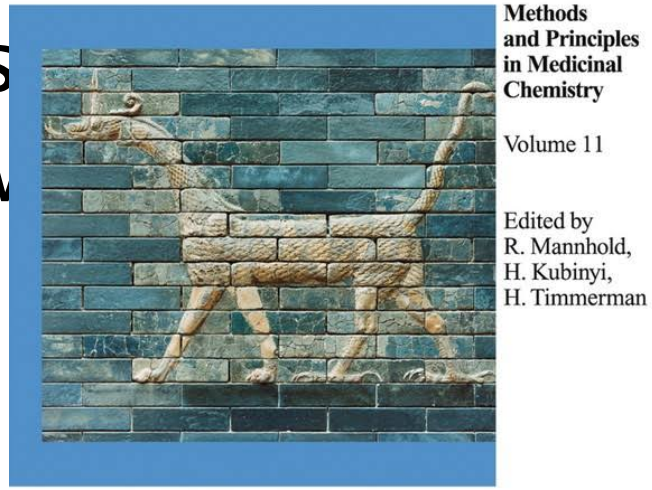
	1	2	3	4	5	s_i
1	0	1	2	2	2	7
2	1	0	1	1	1	4
3	2	1	0	2	2	7
4	2	1	2	0	2	7
5	2	1	2	2	0	7

$$W = \frac{1}{2}(7 + 4 + 7 + 7 + 7) = 16$$

$$R = 2.00$$

A large variety of molecular descriptors

- Over 3000 molecular d
  defined
- Electronic, topology, g
  bonding, substituent..
  influence different des
  - Different points of view
    molecule



Some software tools

- ChemOffice suite
- Marvin Suite
- Chemicalize on line site
  - https://chemicalize.com/
- E-Dragon on line software
  - http://www.vcclab.org/lab/edragon/
- R environment packages QSARdata

Statistical models

- Linear models
  - Multilinear regression (MLR)
    - Problem $p>>n$: a lot more parameters (p) than data (n), collinearities between parameters
  - Dimension reduction
    - A restrained number of linearly independent "components", linear combinations of the initial parameters, is used
      - Regression PCR (*Principal Component Regression*)
      - Regression PLS (*Partial Least Squares*)
- Non-linear models
  - Neural networks, Random forests
  - Bayesian classification

PCR and PLS

- PCR
  - We compute principal components along the axes that maximise the variance, then we model the regression of Y on these components
- PLS
  - We compute components that maximise the covariance with the response Y

An application example

"Comparison between 5,10,15,20-Tetraaryl- and 5,15-Diarylporphyrins as Photosensitizers: Synthesis, Photodynamic Activity, and Quantitative Structure-Activity Relationship Modeling" *J. Med. Chem.* (2006)

(Published by the group of Pr P. Gramatica, University of Insubria, Italy)

$$\log 1/IC50 = -10.06 + 21.05GATS6v - 40.55PW3 + 36.15R4u$$

- Model adjusted on data from 34 molecules:
  - 22 molecules used as training set (model fitting)
  - 12 molecules used as validation set
- A genetic algorithm used to select the adequate descriptors for MLR
- The descriptors selected express size and shape features of the molecules
- The model was used to predict IC50 for five more molecules, to select which were more interesting to test

Some software tools (and a conference)

- R environment package DEMOVA
  - Also packages PLS, FactoMineR, spls…
- On-line VCCLAB tools
  - ASNN, PNN, UFS, PLS
- And a lot of commercial softwares:
  - Accelrys, Sybyl,
- Also check:
  - https://www.qsartoolbox.org/home
  - http://www.euroqsar2018.org/
    - 22nd EuroQSAR Thessaloniki, September 16-20, 2018