Laboratoire ITODYS, CNRS - UMR 7086
Université Paris 7 – Denis Diderot

MolDIA

XML Based System for Virtual Screening Using a Fragment Approach

By: Ana Maldonado

Workshop Chemoinformatics in Europe: Research and Teaching, 29 May – 1 June 2006

Outline

- Problematic
- Concepts overview
  - XML for Structuring Information
  - Similarity and Diversity
- MolDiA Software
  - Molecular Databases
  - Descriptors and Similarity Computation
- Results of MolDiA using ZINC
- Conclusion
- Future Work

Problematic

- Management of chemical databases:

  - bigger, more complex, no metadata...

  => How to structure the data, extract information & knowledge?

- Search for new molecules:

  - more similar (activity), more diverse (structure)

  => How to reconcile both views?

Structuring Chemical Information

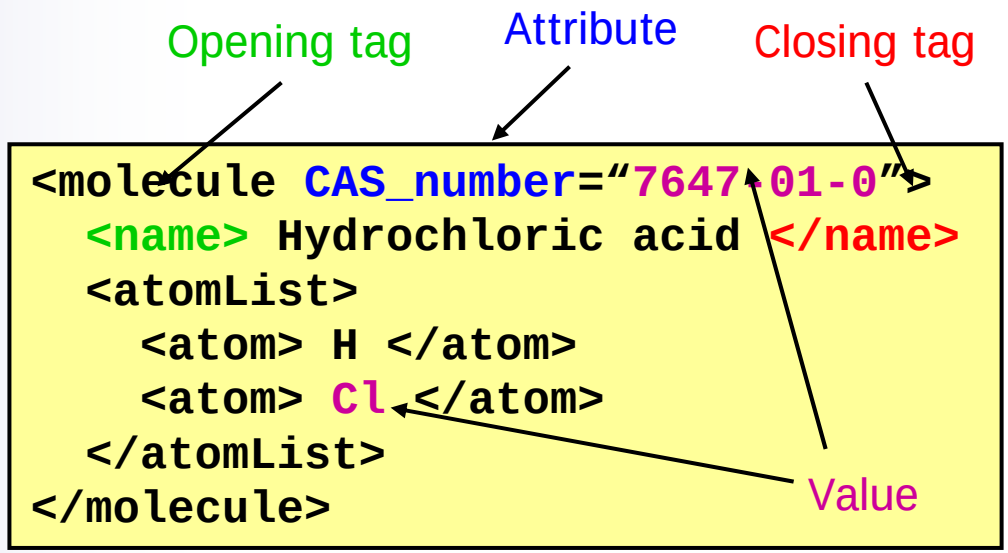
- XML (Extended Markup Language):

- A set of named tags. Each opening tag has its matching closing tag in order to form a tree
- A set of attributes / values for each tag
- Rules for controlling the ordering & nesting of the tags (DTDs)
- A markup language for Chemistry: CML

- How I use XML in Chemistry?

- Other applications:

ThermoML, SpectroML, GAML, MedML



Similarity and Diversity

- Fuzzy and relative concept:

  $\Rightarrow$ we need quantitative definitions

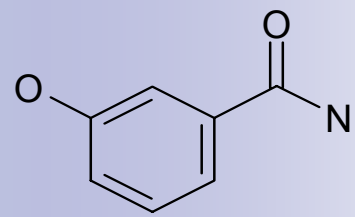
- In chemistry:
  - Aristote's *Scientific method*
  - Mendeleev's *Periodic Table*
  - *Similar property principle*

  $\Rightarrow$ similar structures $\rightarrow$ similar properties

- Similarity and Diversity are complementary

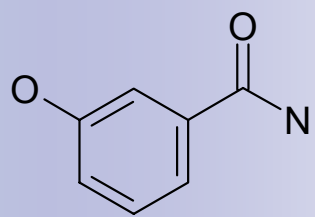
MolDiA: Molecular Diversity Analysis

- Main components:
  - Molecular Databases → FragDB, CompDB, QueryDB
  - Molecular Representations → descriptors

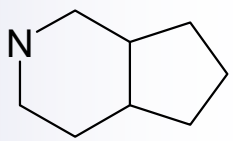


<DescriptorVector>

  - Similarity Computation → indices

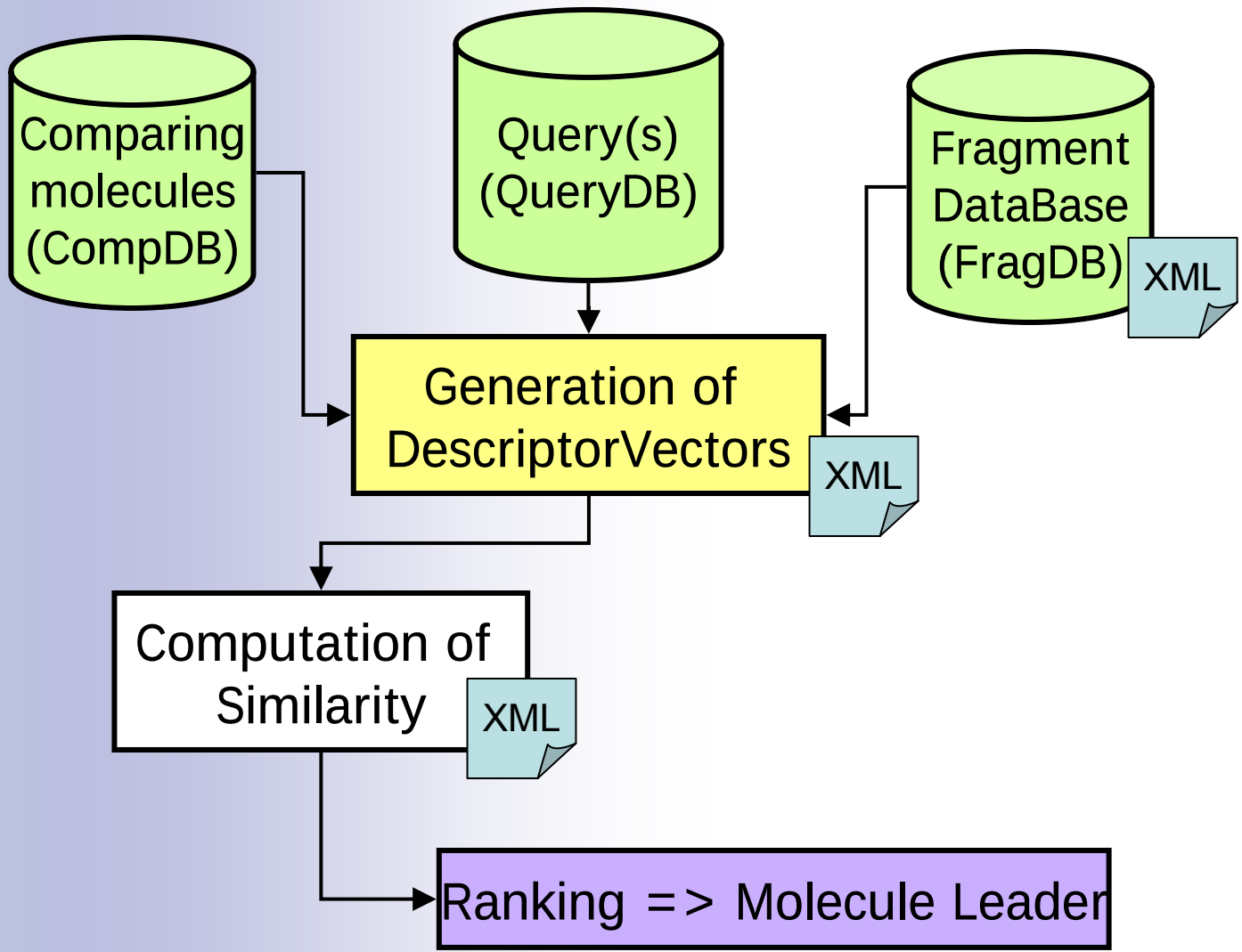


vs

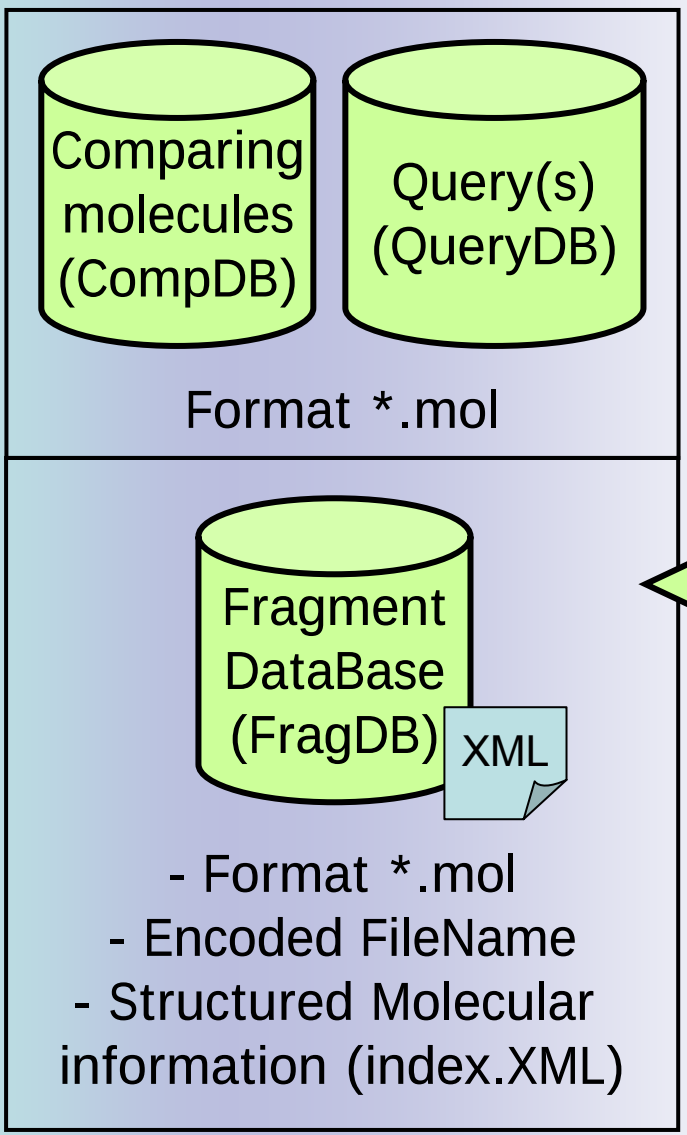


Tanimoto, Simpson, Cosinus ...

MolDiA Overview



Molecular Databases




```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<index>
  <File name = "CAUN5-156Qb.mol">
    <Keys>
      <Key name = "FID" value = "156Qb"/>
      <Key name = "FAtomSum" value = "8"/>
      <Key name = "FRing" value = "5"/>
      <Key name = "FGF" value = "none"/>
    </Keys>
    <Properties>
      <Property name = "HBondAD" value = "1"/>
      <Property name = "PotPCharged" value = "1"/>
      <Property name = "PotNCharged" value = "0"/>
      <Property name = "HydPhi" value = "1"/>
      <Property name = "Aromat" value = "1"/>
      <Property name = "Polar" value = "1"/>
      <Property name = "HydPho" value = "0"/>
    </Properties>
  </File>
  ....
</index>
```

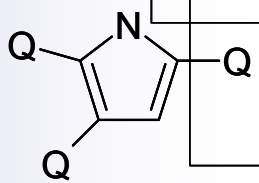
C = Cyclic A = Acyclic

A = Aromatic, N = Non_Aromatic, G = Group_Functional

CAUN5...

U = Single_Cycle, T = Fused_Triple_Cycle, S = Saturated
D = Fused_Double_Cycle, Q = 4+Fused_Cycle, I = Unsaturated
C= Carbonyl

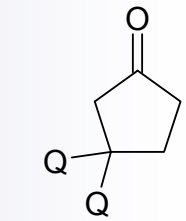
C = Carbon_atom, O = Heteroatom_O, **N = Heteroatom_N**
S = Heteroatom_S, M = Mixture_Heteroatoms, W = Special_atom
X = Halogen_atom



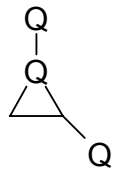
3-9 = Number_atoms1ring, 3-9 = Number_atoms3ring
3-9 = Number_atoms2ring 3-9 = Number_atoms4ring, etc.…

Composition of FragDB

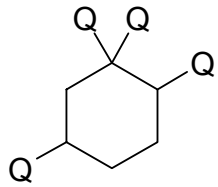
Some examples…



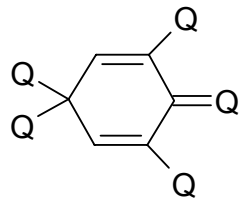
CNUO5-105b



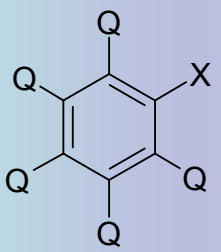
CNUQ3-131f



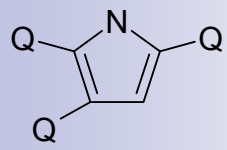
CNUQ6-074bi



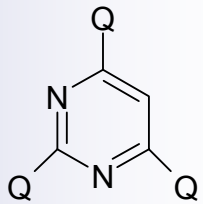
CNUQ6-169u



CAUX6-055X



CAUN5-156Qb

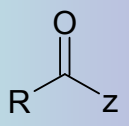


CAUN6-153Qc

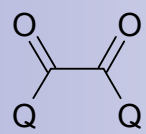
502 Frag CYC
61 Frag ACYC
321 Rules

Fragment DataBase (FragDB)

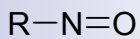
A H Q
R Z X M



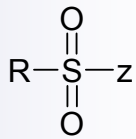
AGCZ-014Z



AGCQ-014Q



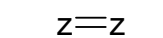
AGIE-038R



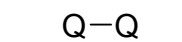
AGIS-051Z

ANIC-003R

ANSX-000X



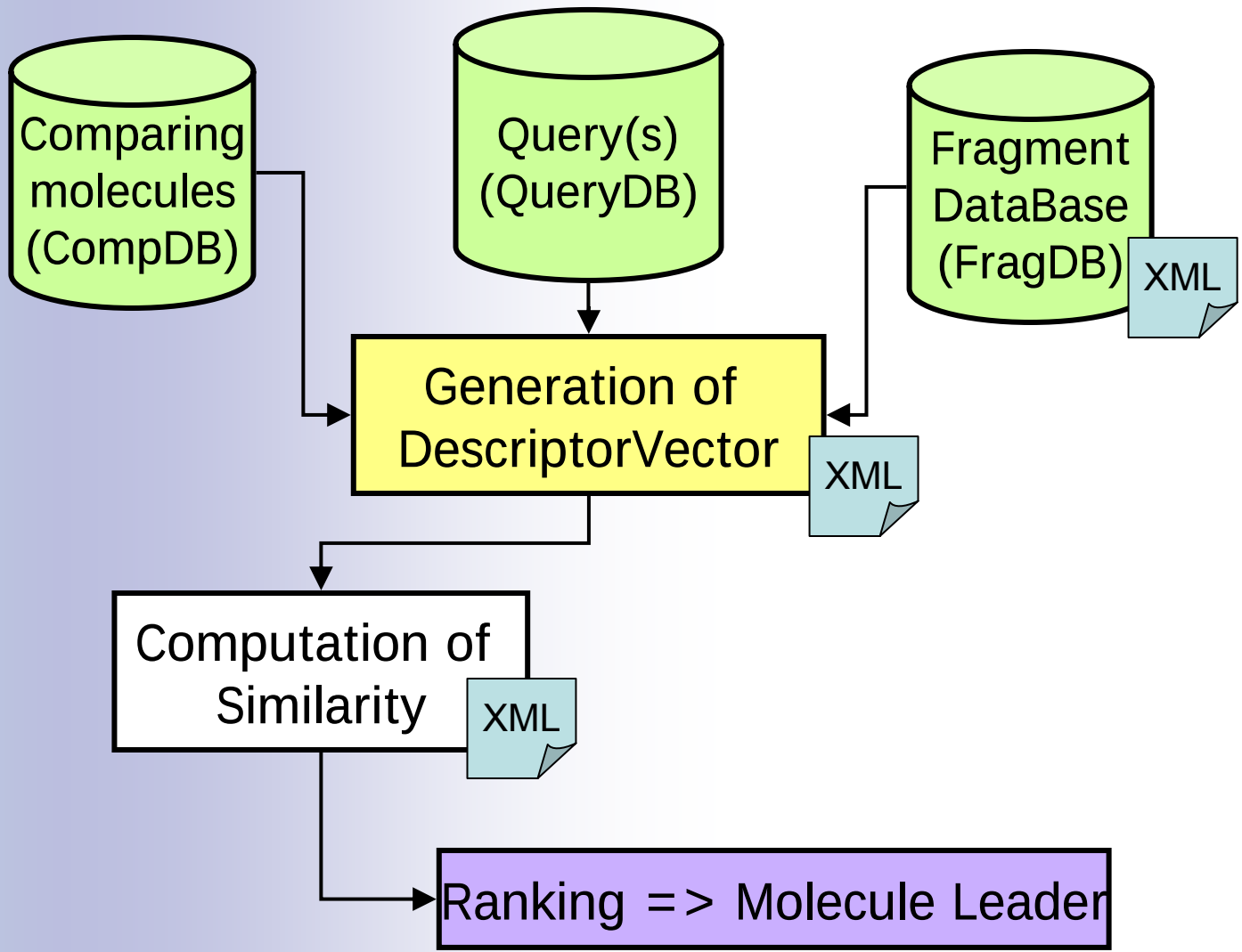
ANIZ-001Z



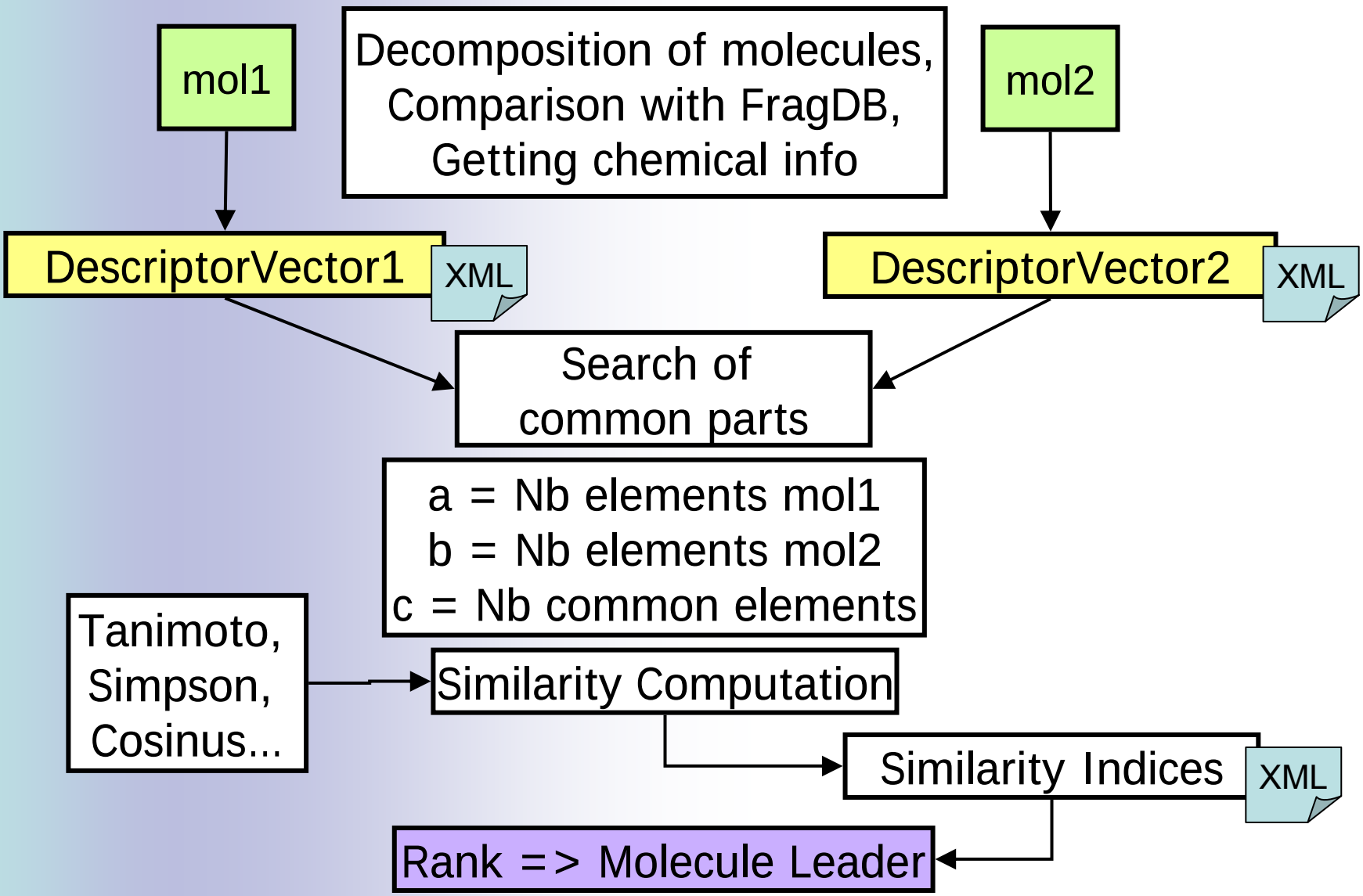
ANSQ-000Q

* Stobaugh, R.E., *J. Chem. Inf. Comp. Sci.* 28 (1988) 180-187
* Ertl, P., *J. Chem. Inf. Comp. Sci*. 43 (2003) 374-380
* Robert P. Sheridan, *J. Chem. Inf. Comput. Sci.* 42 (2002) 103-108
* Xu, J., Stevenson, J., *J. Chem. Inf. Comput. Sci.* 40 (2000) 1177-1187

MolDiA Overview

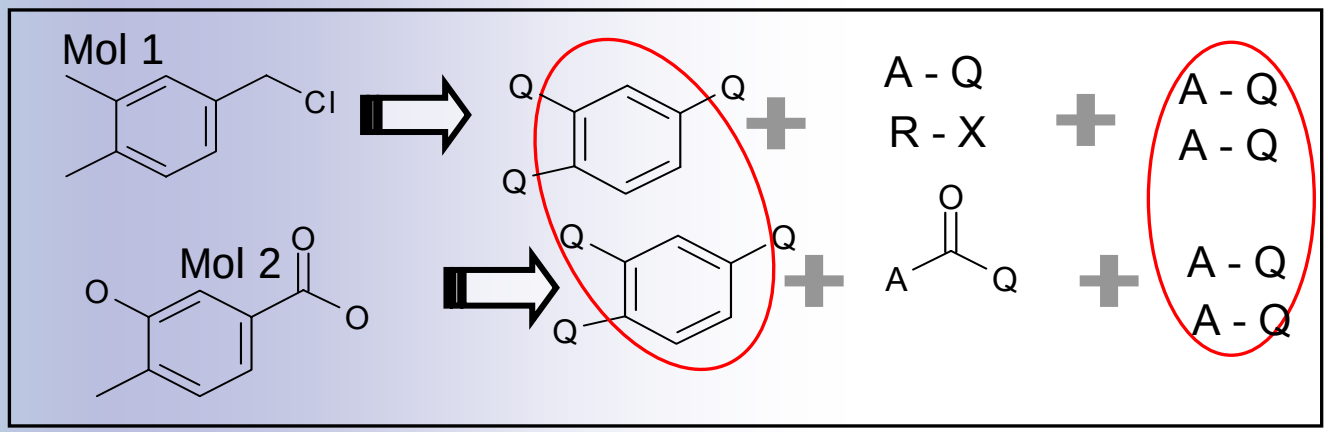


Descriptor and Similarity Computation



Descriptor and Similarity Measures

- Descriptors: Structural vectors



Mol 1: <CAUC6-054a, ANSA-000X, ANSA-000Q, ANSA-000Q, ANSA-000Q >

Mol 2: <CAUC6-054a, AGCQ-014Q, ANSA-000Q, ANSA-000Q >

- Multiple indices $\Rightarrow$ Multiple measures

Tanimoto St=c/(a+b-c)
Cosinus Sc=c/$(ab)^{1/2}$
Simpson Ss=c/min(a,b)

a = number of fragments mol 1
b = number of fragments mol 2
c = common fragments

a	b	c
5	4	3

St=0.50
Sc=0.67
Ss=0.75

At least 27 possibilities of Analysis...

- With different levels of comparison:
  - structural information only,
  - or with physicochemical properties,
  - or with structural / property weights...
- Different Similarity and Diversity analysis:

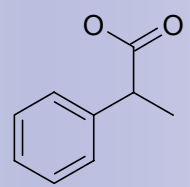
  1-1, 1-N, N-N, N-M...
- Several measures of Similarity:

  Tanimoto, Simpson, Cosinus...

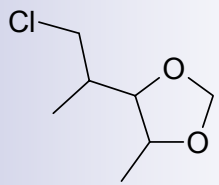
=> users can customize the computation

Results of MolDiA using ZINC DB (I)

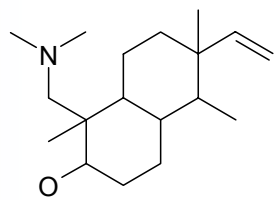
- CompDB: 40 molecules from ZINC - A free database for virtual screening, http://www.blaster.docking.org/zinc/
- Indices used: Tanimoto, Simpson, Cosinus
- No weight customization
- Kind of analysis: 1-N and N-N
- QueryDB for 1-N: 4 query molecules, for N-N: all!



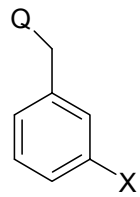
Query1Z2.mol,



Query2Z2.mol,



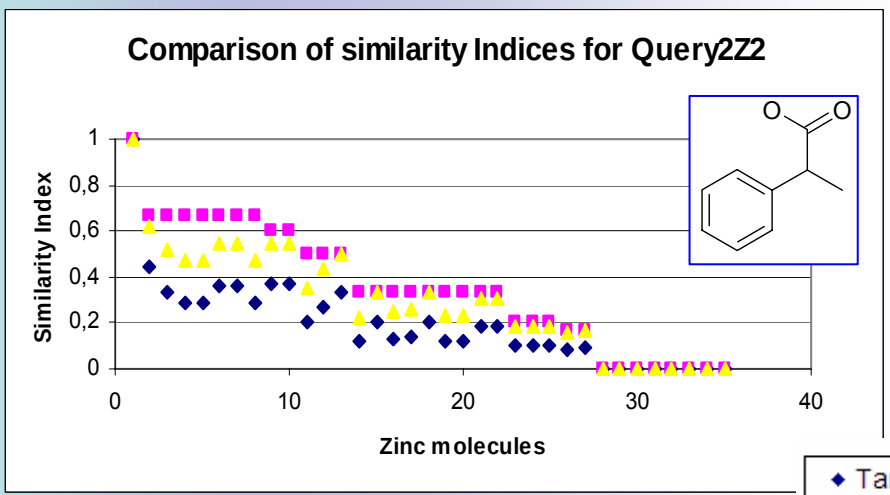
Query1Z3.mol,

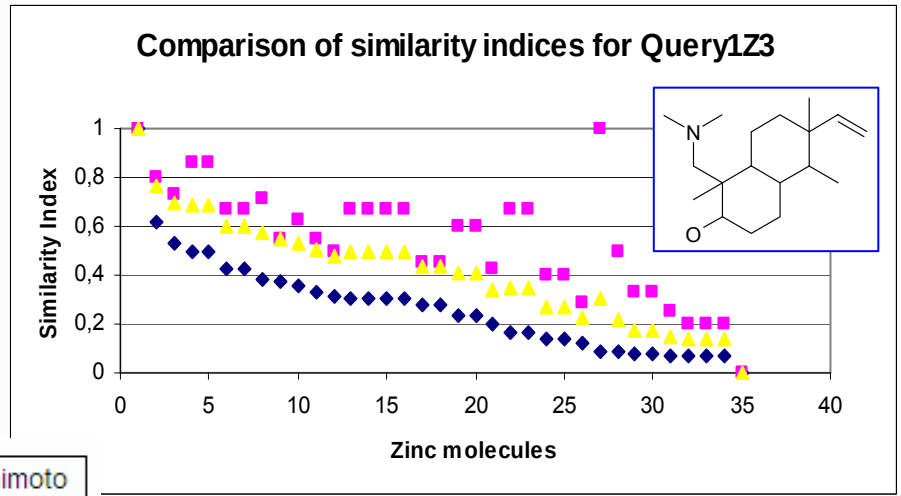


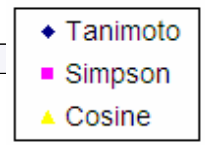
QueryGen.mol

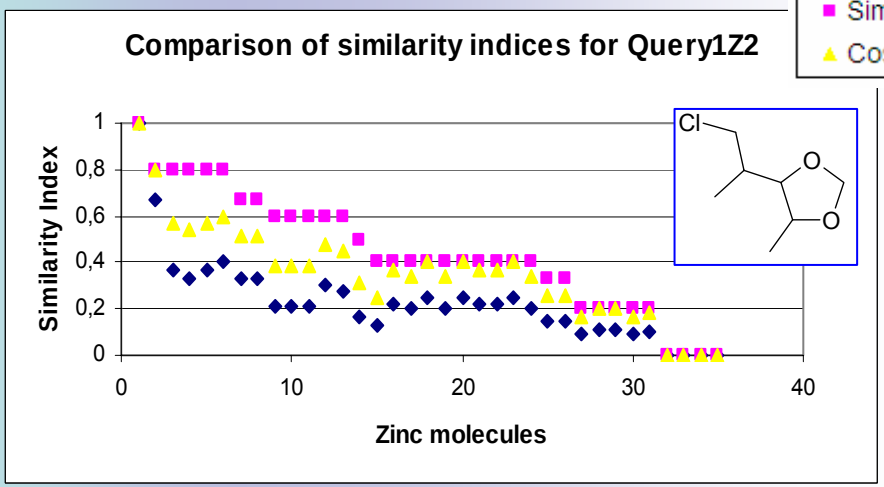
Results of MolDiA using ZINC DB (II)

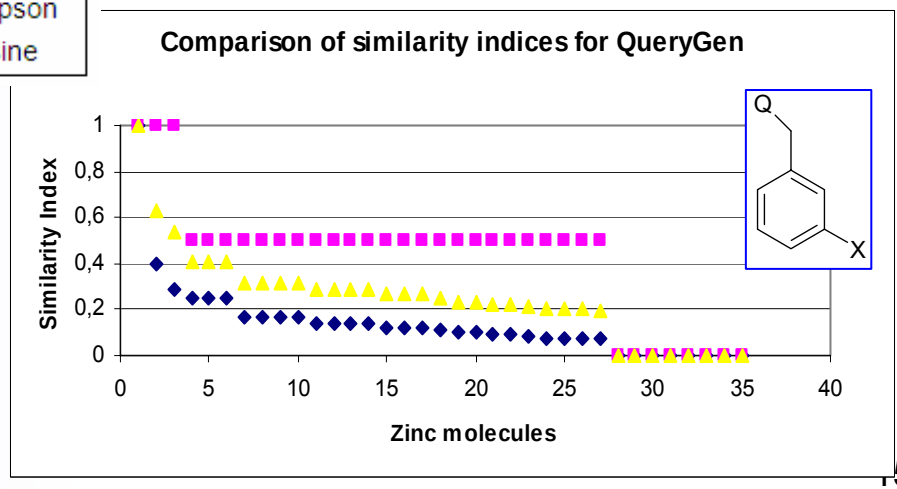
- Ranking of similarity indices for an 1-N analysis of QueryDB molecules:





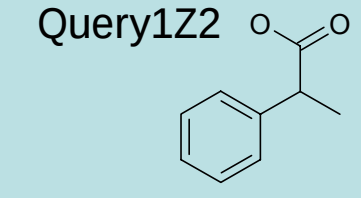
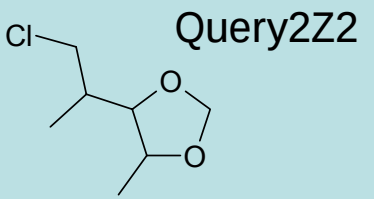
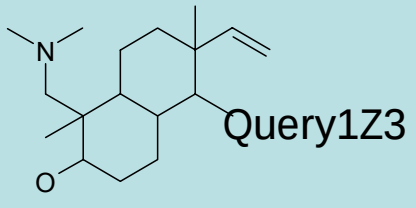
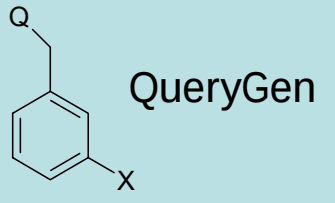
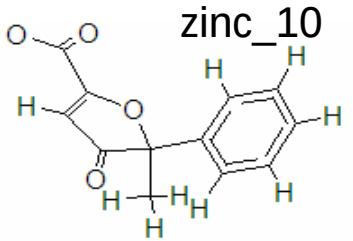
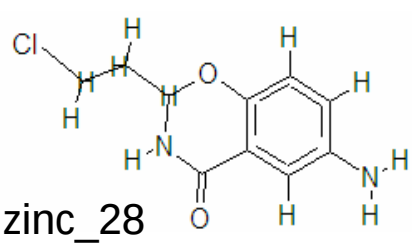
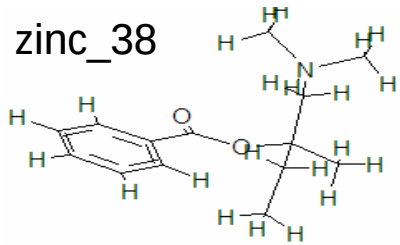
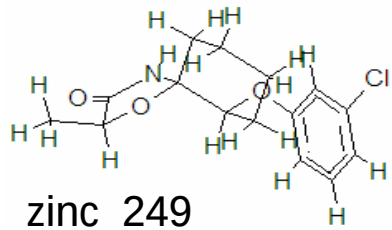






Results of MolDiA using ZINC DB (III)

- Some computed Similarity values....

mol1 (Q)	mol2	Tanimoto	Cosinus	Simpson
	Query1Z2	0.66	0.8	0.8
	Query2Z2	0.38	0.55	0.6
	Query1Z3	0.62	0.76	0.8
	QueryGen	0.29	0.53	1
	zinc_10			
	zinc_28			
	zinc_38			
	zinc_249			

- Number and percent of molecules with similarity value ≥ 0.8

	Measure of Sim ≥ 0.8							
	Query1Z2		Query2Z2		Query1Z3		QueryGen	
Tanimoto	1	2.94%	1	2.94%	1	2.94%	1	2.94%
Cosinus	2	5.88%	1	2.94%	1	2.94%	1	2.94%
Simpson	6	17.65%	1	2.94%	5	14.7%	3	8.82%

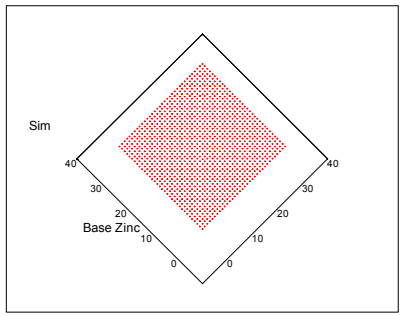
Results of MolDiA using ZINC DB (III)

- Diversity N-N analysis for Zinc DB

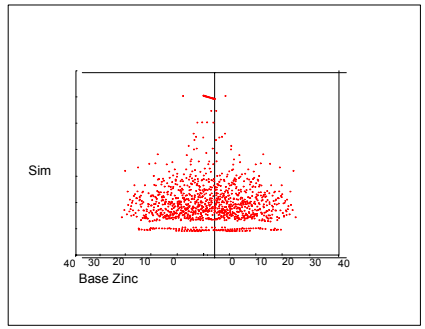
a) Using Tanimoto

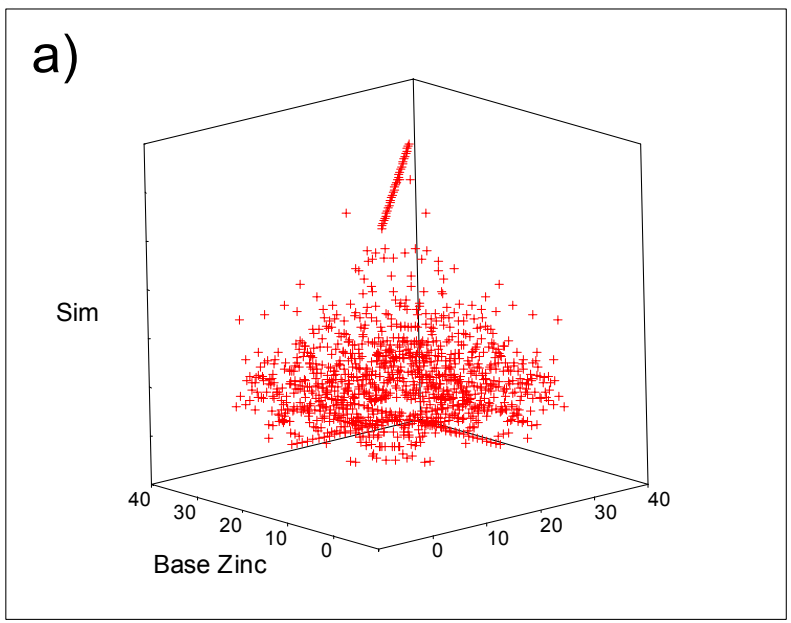
b) Using Simpson

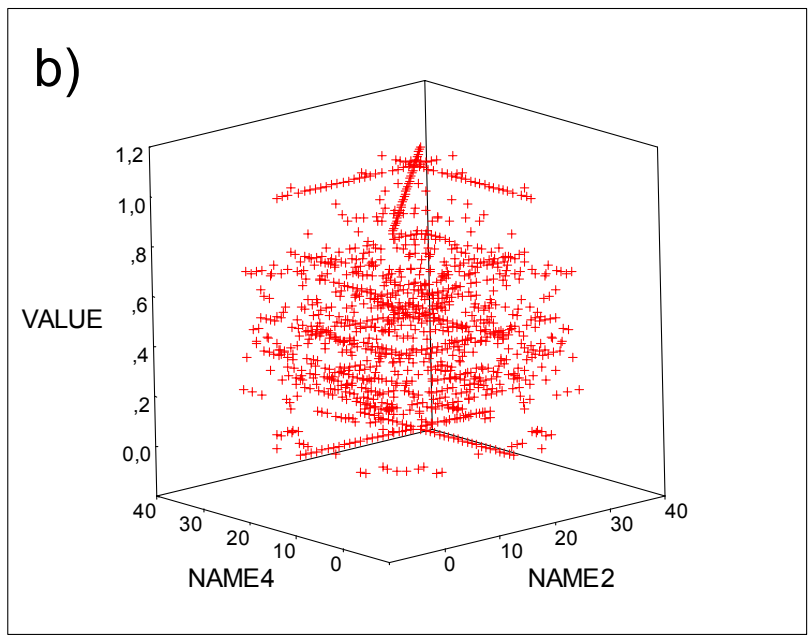
c) Symmetric matrix => symmetric plots



c)







DEMONSTRATION

Conclusion

- Virtual screening based on a new extension of the diversity concept for the drug design => economic and fast research of molecules
- Use of generic atoms and fuzzy atomic comparison for fragment matching on queries => large flexibility in the search process
- Use of Markup Language (XML) in MolDiA => structure, process and exchange complex chemical data, better compatibility with the Web. Possibility to add a stylesheet for textual output in the Web
- Different levels of comparison and use of different kinds of weights => customization of the analysis
- Use of different similarity measures => possibility to effectuate data fusion

Future work

Mid-term perspectives

- Implementation of a similarity/diversity formula editor
- Implementation of a graphic module for drawing query or test molecules
- Extension of the FragDB
- Extension of physicochemical properties

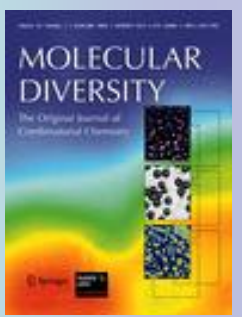
Long term perspectives

- Design and implementation of a **QSAR** module
- Extension of functionalities for application in molecular **bio**logy and **bio**informatics
- Similarity/Diversity analysis for **3D** molecules
- **Molecular information** structured using CML

Some references

Ana Maldonado, Michel Petitjean, Jean-Pierre Doucet, Annick Panaye and Bo Tao Fan. MolDIA: XML based system of molecular diversity analysis towards virtual screening and QSPR.*SAR and QSAR in Environmental Research* 17(1): 11-23 (2006)

Ana Maldonado, Using XML for Structuring the Chemical Information: Towards a Chemical Knowledge Representation. Published by *MDPI Online Edition*. ISBN 3-906980-17-0 (2005) http://www.mdpi.org/fis2005/proceedings.html



Ana Maldonado, Michel Petitjean, Jean-Pierre Doucet and Bo Tao Fan. Molecular Similarity and Diversity: Concepts and Applications. Review article, *Molecular Diversity*, 10(1): 39-79 (2006)

CMTPI-2005

QSAR 2006

Thanks for your attention!

Questions?

Laboratoire ITODYS, CNRS - UMR 7086
Université Paris 7 – Denis Diderot