MolDIA: Fragment approaching to Molecular Similarity and Diversity Analysis

Ana G. Maldonado, Michel Petitjean, Jean-Pierre Doucet, Annick Panaye and Bo Tao Fan*

ITODYS, Institut de Topologie et de Dynamique des Systèmes, CNRS UMR-7086, University Paris-7,
1, rue Guy de la Brosse, 75005 Paris, France

* Corresponding author email : fan@paris7.jussieu.fr

Objectives & Motivation

Management of chemical DataBases:

- Modern DB: increasing size and complexity with time

=> How to structure the data?

Search of new molecules:

- More similar (for the activity) AND more diverse (for the structure)

=> How to reconcile both views?

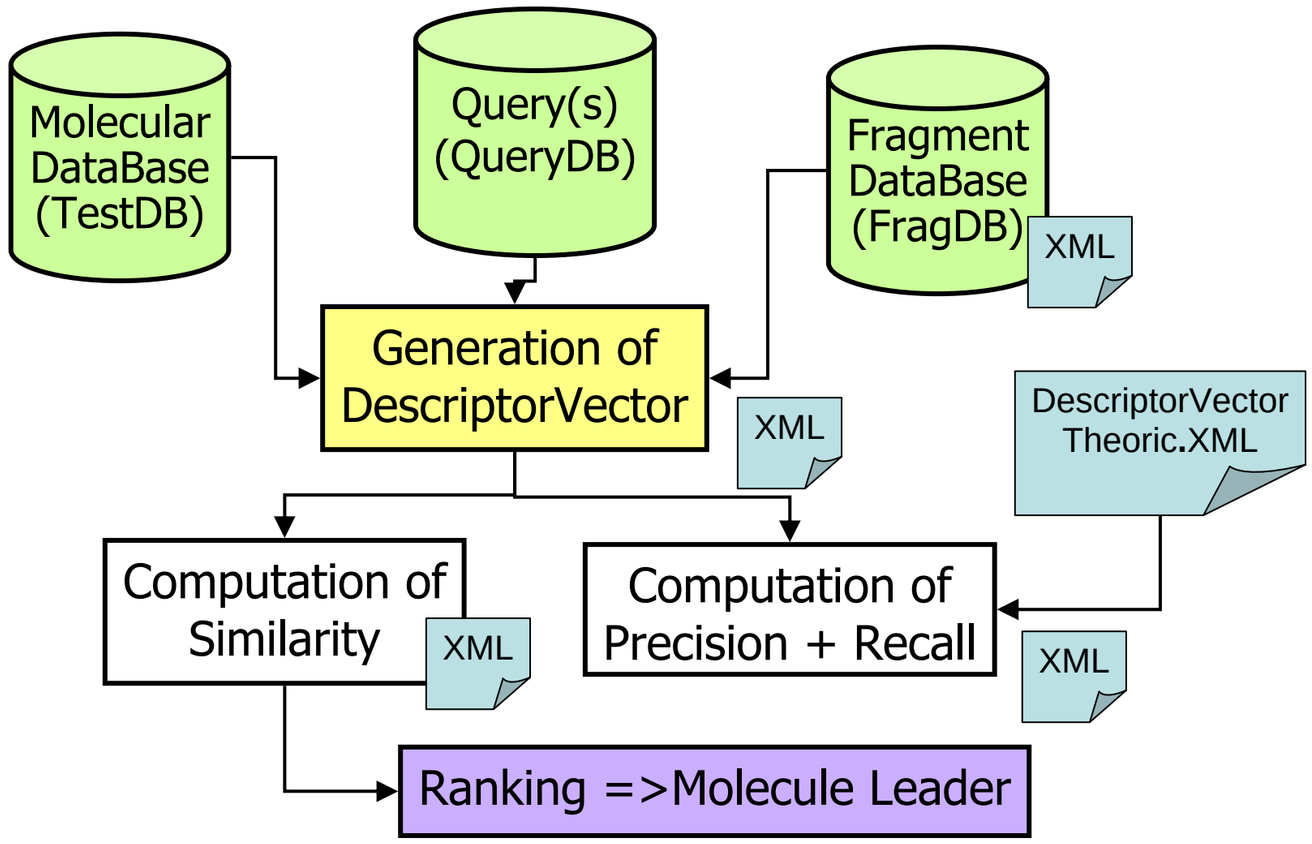
MolDiA (Molecular Diversity Analysis)

Main components:

1) Molecular Databases
-> FragDB, TestDB, QueryDB

2) Molecular Representations
-> <DescriptorVector>

3) **Similarity Computation**
-> Different indices



MolDiA Molecular Databases

Molecular DataBase (TestDB)

Query(s) (QueryDB)

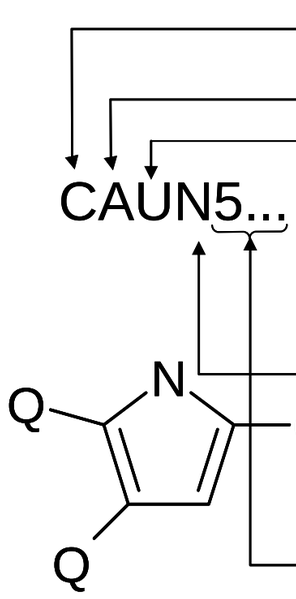
- Format : .mol File
- Molecular information analysed as given by the user

502 Frag CYC
61 Frag ACYC
321 Rules

Fragment DataBase (FragDB)

XML

- Format .mol
- Encoded FileName
- Structured Molecular information (index.XML)
- Exclusion rules defined



C = Cyclic A = Acyclic

A = Aromatic, N = Non_Aromatic, G = Group_Functional

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

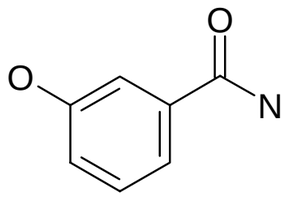
3-9 = Number_atoms1ring, 3-9 = Number_atoms3ring
3-9 = Number_atoms2ring 3-9 = Number_atoms4ring, etc.…

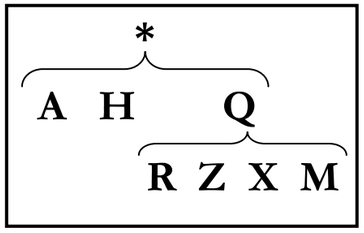
```
<?xml version="1.0" encoding="iso-8859-1" standalone="no" ?>
<index>
  <File name = "CAUN5-156Qb.mol">
    <Keys>
      <Key name = "FID" value = "156Qb"/>
      <Key name = "FAtomSum" value = "8"/>
      <Key name = "FRing" value = "5"/>
      <Key name = "FGF" value = "none"/>
    </Keys>
    <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>
```

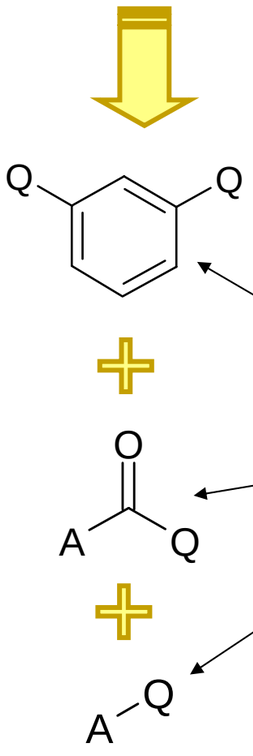
Descriptors and Similarity Measures

Descriptors: Structural vectors

- Substructural/Fragmental approach
- Use of generic atoms
- Example: 3-Hydroxy-benzamide







<DescriptorVector>:

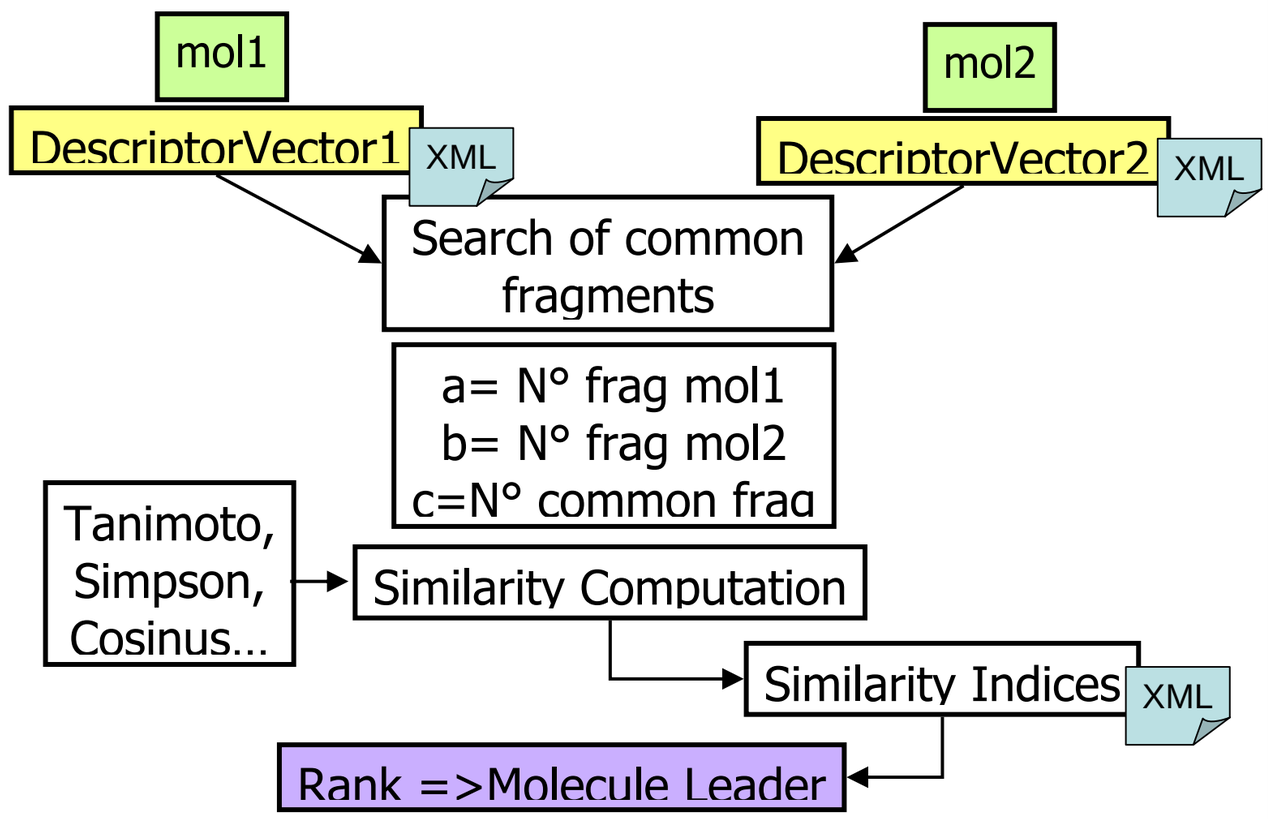
< 3-Hydroxy-benzamide;
<CAUC6-54i,
AGCO-015A,
ANSA-000A >

Multiple indices ⇒ Multiple measures

- With different levels of comparison:

Structural information only and/or weights for the structures & physicochemical properties

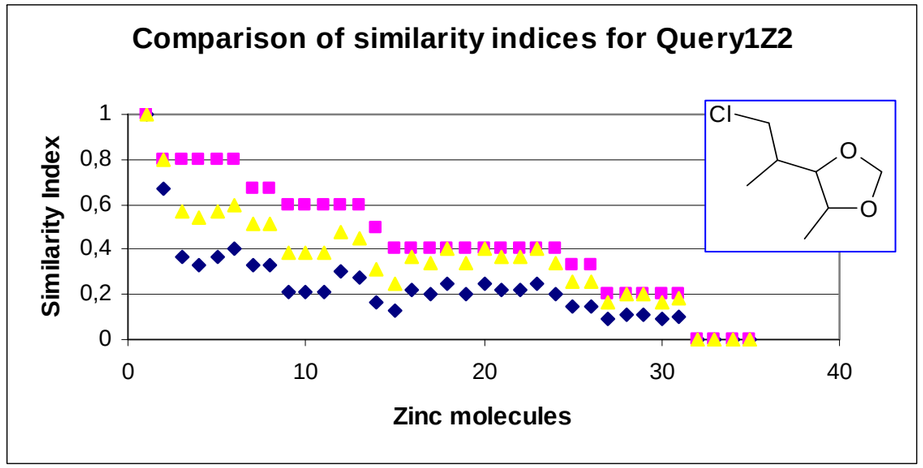
- Different analysis: 1-1, 1-N, N-N, N-M…
- Several measures of Similarity: Tanimoto, Simpson, Cosinus…

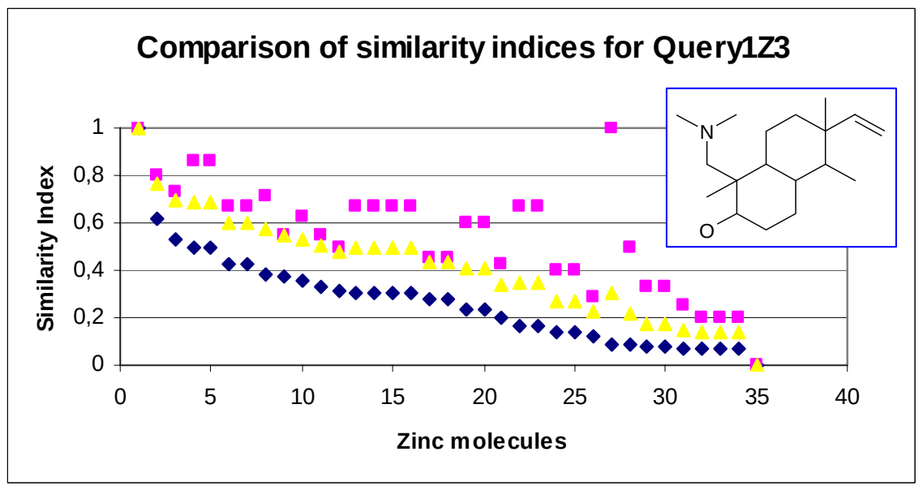


Results of MolDIA using ZINC (I)

Comparison of similarity indices

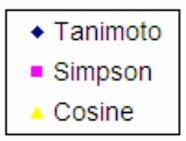
- TestDB: ZINC- A free database for virtual screening http://www.blaster.docking.org/zinc/
- Indices used: Tanimoto, Simpson, Cosinus
- No customization of weights
- Kind of analysis: 1-N (one molecule Vs one DB)
- QueryDB: 3 query molecules

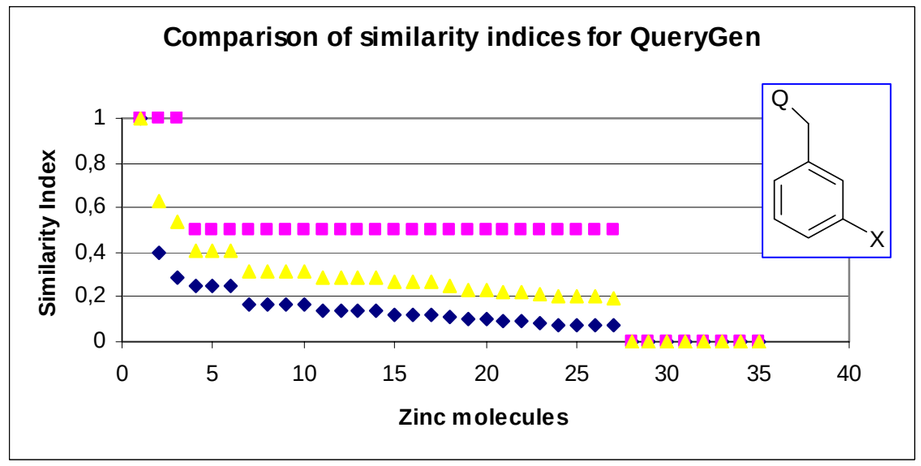




Number and percentage of molecules with similarity value ≥ 0.8

	Measure of Sim ≥ 0,8					
	Query1Z2		Query1Z3		QueryGen	
Tanimoto	1	2,94%	1	2,94%	1	2,94%
Cosinus	2	5,88%	1	2,94%	1	2,94%
Simpson	6	17,65%	5	14,7%	3	8,82%



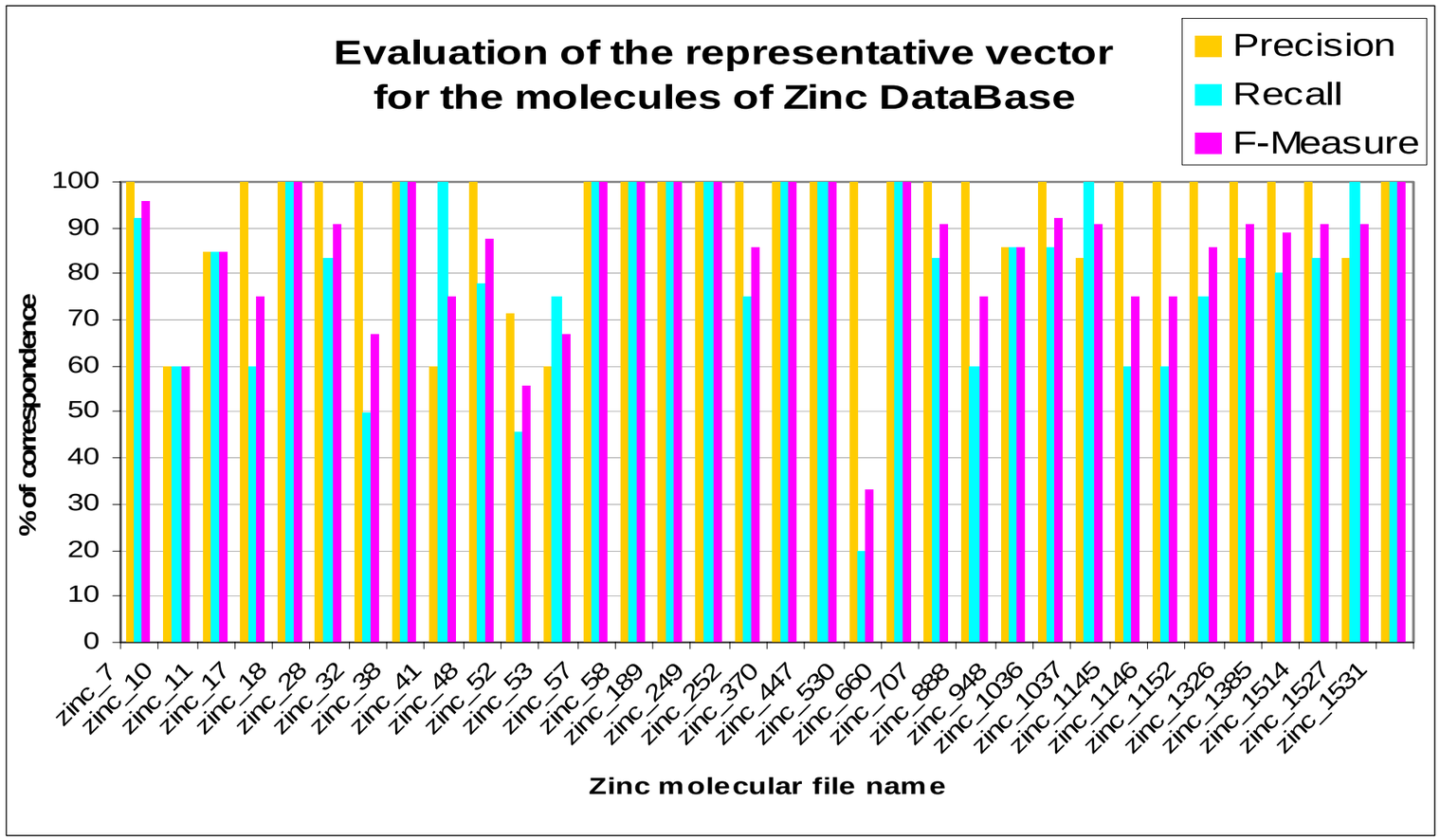


Results of MolDIA using ZINC (II)

Evaluation of <DescriptorVector>

- Computation of Precision, Recall and F-measure
- % of correspondence between theory and experimental structures

	Number of molecules with 100% of correspondence		Number of molecules with 80% of correspondence		Number of molecules with less than 50% correspondence	
Precision	26	76,47%	30	88,24%	0	0%
Recall	10	29,41%	22	64,71%	3	8,82%
F-Measure	10	29,41%	24	70,59%	2	5,88%



$$\text{Precision} = \frac{Vt \cap Vg}{Vg}$$

$$\text{Recall} = \frac{Vt \cap Vg}{Vt}$$

$$\text{F-measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

- Vg = <DescriptorVector> generated by the software
- Vt = <DescriptorVector> constructed by human (theory)

Conclusions & Future work

Conclusions

- MolDiA: a virtual screening application using a new extension of the diversity concept for drug design
- Use of generic atoms in the fragments AND in the queries => large flexibility in the search process
- Markup Language (XML) in MolDiA => structure, process and exchange complex chemical data, better compatibility with the WEB
- Different levels of comparison & use of several kinds of weights => customize the computation
- Use of different similarity measures => data fusion techniques

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

Long term perspectives

- Design and implementation of a QSAR module
- Extension of functionalities for application in molecular biology and bioinformatics
- Similarity/Diversity analysis for 3D molecules
- Extension of physicochemical properties

Bibliography

A.G. Maldonado, M. Petitjean, J-P. Doucet, A. Panaye and B.T. 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. Presented at the Third International Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2005). Oct 29th to Nov 1st. Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences. Shanghai,China.