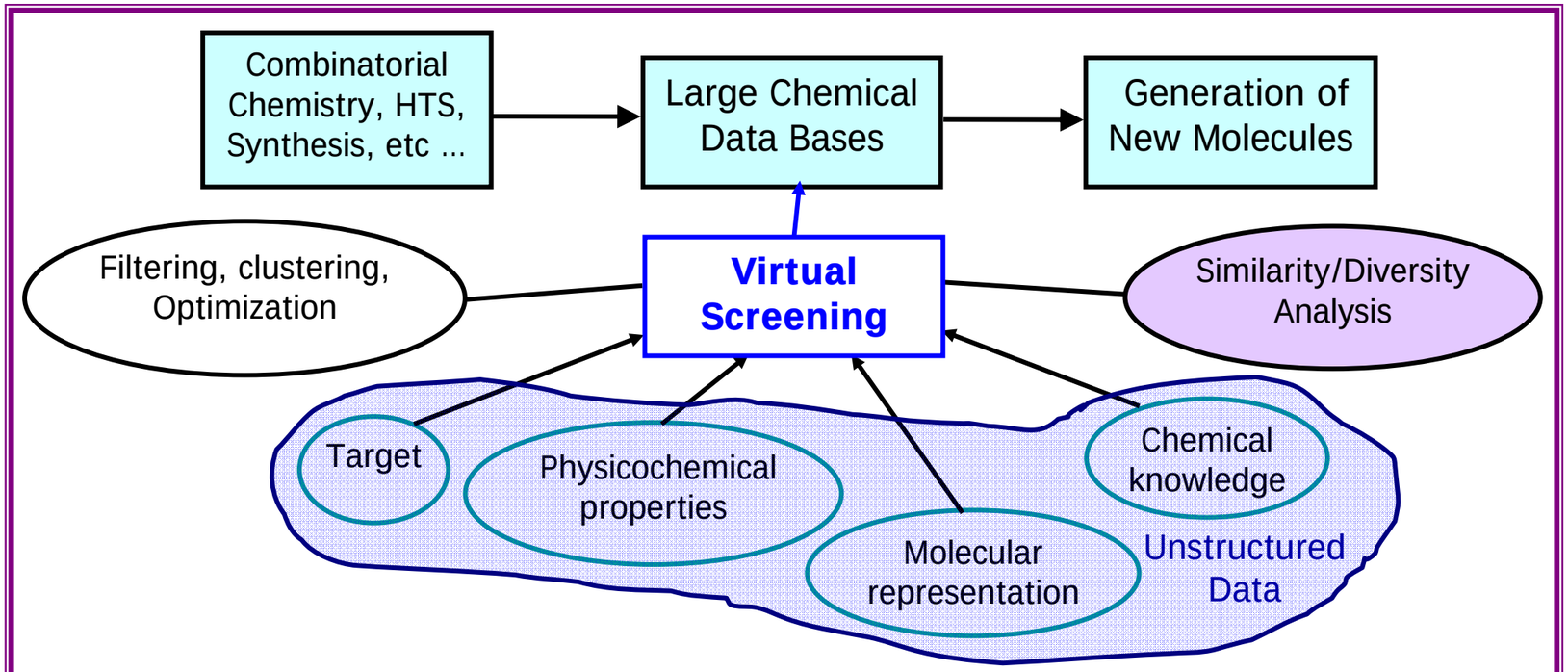
MolDIA: XML Based System of Molecular Diversity Analysis Towards Virtual Screening and QSPR

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

INTRODUCTION

Since a few decades, new techniques have appeared to enrich the pharmaceutical "chemical panorama" and to address new sources of **molecular diversity**. These techniques have allowed medicinal chemists to improve drug discovery by generating large (virtual or real) chemical databases. The problem is that the database size increases dramatically with time, and now, the volume of the chemical information available is huge. Screening and data-mining techniques improve the analysis, retrieval and management of big quantity of data.



Virtual Screening: from chemical information to new structures

- Screening tools includes:
  - Clustering, filtering, classification, etc.
  - 3D tools (docking, electronic density, fields…)
  - (Sub)structure searching
  - Similarity searching
  ....
- Screening techniques includes:
  - High Throughput Screening (HTS) and
  - Virtual screening (VS)
  ....

In **Virtual Screening,** the data available (e.g. molecular structures, physicochemical properties, etc.) is processed to build models that allow the treatment of these data. Sometimes these data are not structured, thus difficult to deal with. Moreover, techniques such as filtering, clustering and optimization, as well as, **analysis of molecular similarity and diversity** will optimize the search of new molecules with respect to a query or group of queries.

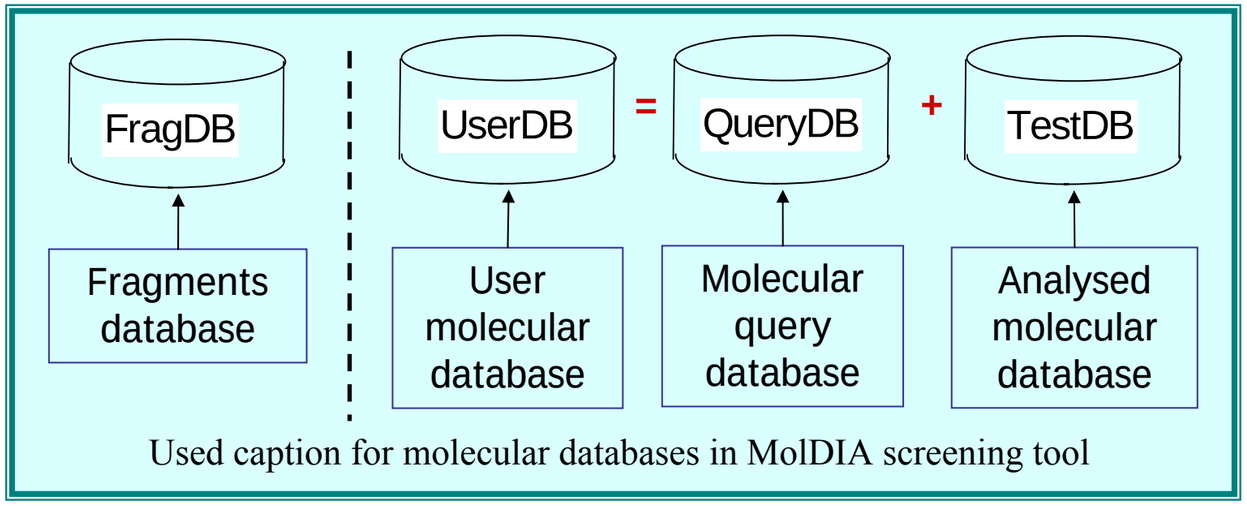
PRINCIPLES

SIMILARITY & DIVERSITY

- Concepts used in chemistry since a long time:
  - Aristote's Scientific Method
  - Mendeleev's periodic table
  - Property Similarity Principle: "molecules with **similar structure** tend to have **similar properties**"
- Their measure has three principal components:
  - The molecular representation → use of descriptors (vectors, equations, numbers, …)
  - The similarity measures → quantitative definitions (indices, coefficients, distances, …)
  - A weighting system → to custom the measures
- The molecular *similarity* provides a simple and popular method for virtual screening and underlies the use of clustering methods on chemical databases. Furthermore, molecular *diversity* analysis explores the way of how molecules cover a given structural space and underlies many approaches for compound selection and design of combinatorial libraries.
- Similarity and Diversity are subjective and fuzzy concepts. Quantitative definitions are necessary if we want to use automatic methods.

MOLECULAR REPRESENTATION

MOLECULAR DATABASES

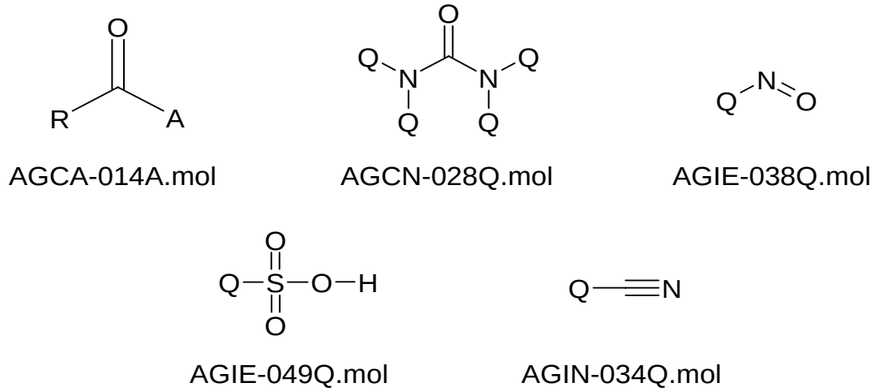


Used caption for molecular databases in MolDIA screening tool

We develop a chemoinformatics tool named **MolDIA** (Molecular Diversity Analisis) dealing with the structure/property relationships. Molecular DBs and particularly the **FragDB** were designed to achieve this comparison. In the **UserDB**, the user selects which molecules will belong to the **QueryDB** and the **TestDB**. All molecules are in MDL-MOL format, and the molecular information is structured using XML.

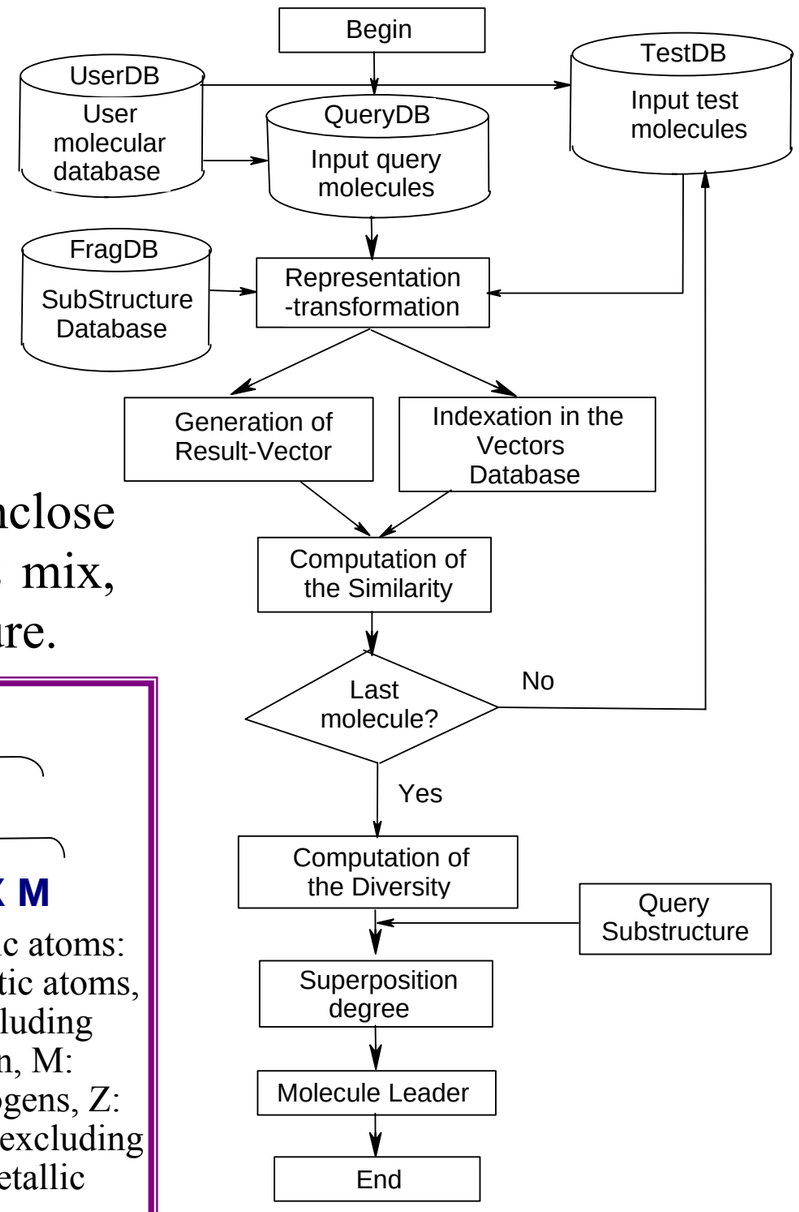
FragDB CONSTRUCTION

The FragDB designed for **MolDIA**, is based on the "property similarity principle", and was constructed using a selection of **functional groups** and popular substructures. Once substructures were chosen, we have used **generic atoms** to multiply the matching frequency without including all molecular single cases.

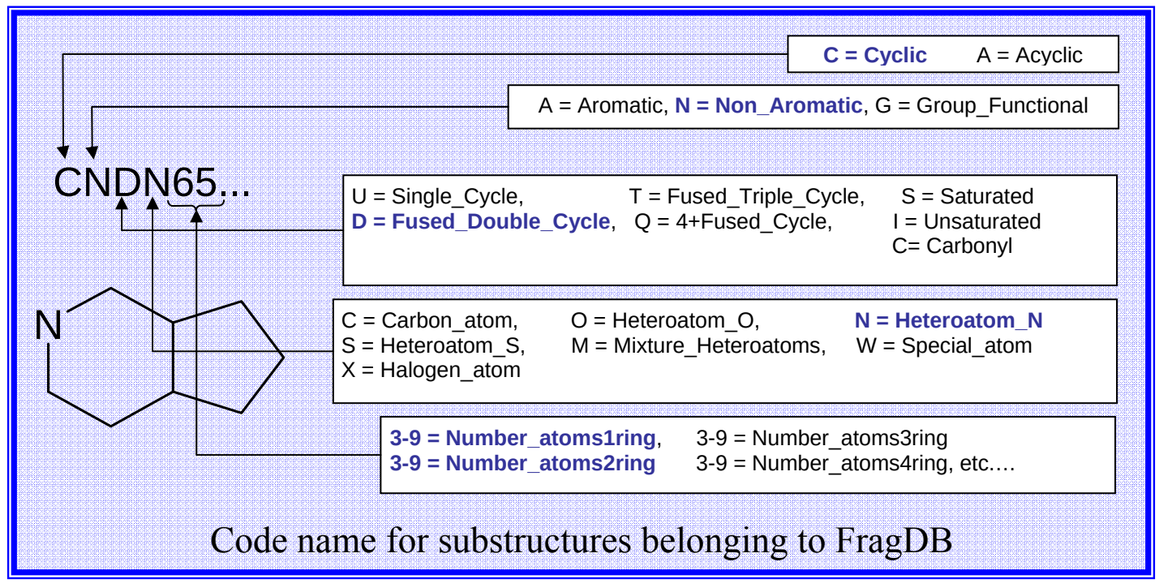


Examples of aliphatic fragments, belonging to FragDB

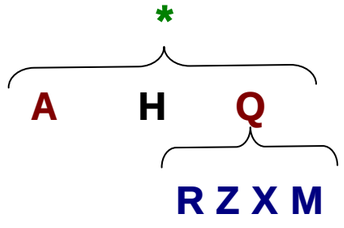
MolDIA Algorithm: First we load the UserDB and FragDB. For each molecule, a descriptor vector is generated using information structured in XML files. Superposition of vector gives the measure of Sim/Div.



When indexing fragments in the DB we use our own **coding system** to enclose fuzzy or un-structured information, as well as, aromaticity, heteroatoms mix, fused cycles, etc. This information will be used to enrich XML file structure.



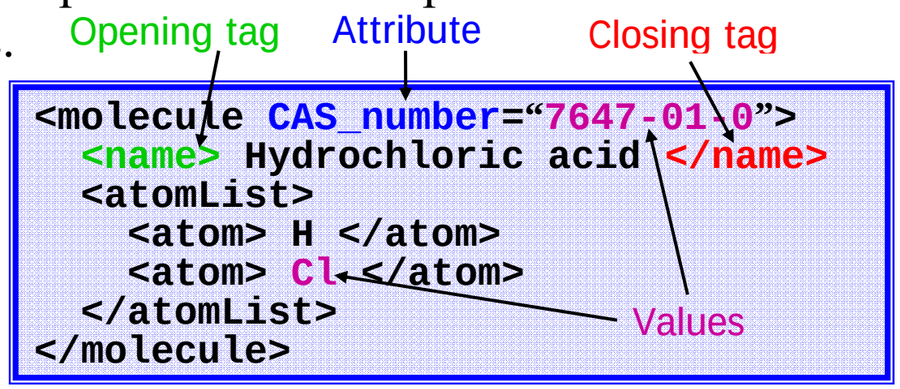
Code name for substructures belonging to FragDB



Classification of generic atoms: *: all atoms, A: aromatic atoms, Q: any heavy atom excluding aromatics and hydrogen, M: metallic atoms, X: halogens, Z: heteroatom B,N,O,P,S excluding C, R: the rest of non metallic atoms.

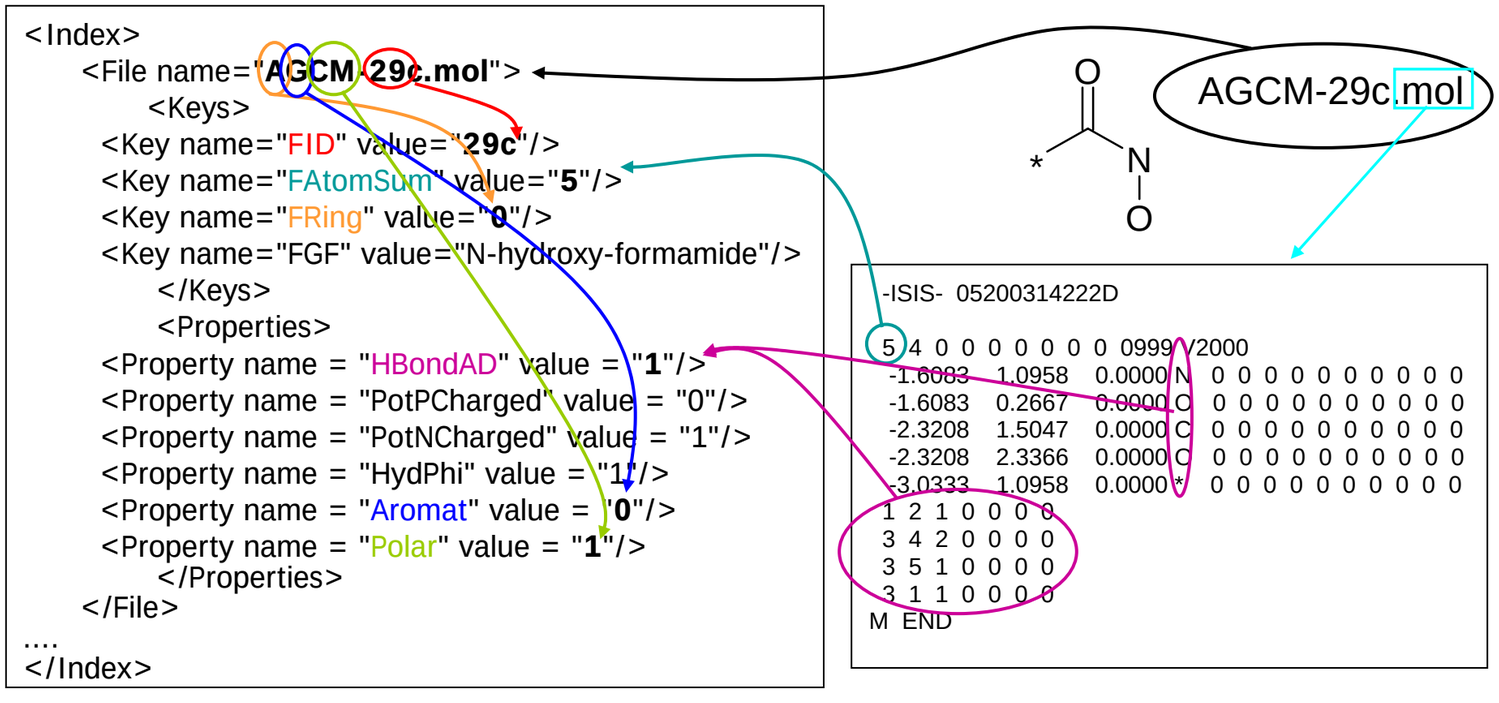
FragDB STRUCTURE

Chemical information systems lacks of a standard structure which allows exchanges in different fields of chemical research. When designing databases for **MolDIA** we encounter this problem. One solution is to use XML, a language for easing the information exchange that has proved to be a powerful alternative to conventional binary file storage and database management systems.



XML consists in a set of named tags, with a particular **syntax**:
- Each opening tag has its matching closing tag (tree structure),
- Each tag has a set of attributes / values,
- There are rules (DTDs) for controlling the ordering and the nesting of the tags ⇒ The data become structured

⇒ The search, management and query of molecular databases in the discovery process are simplified.



Filling of the Index.XML file: the data is first parsed and then transformed into internal data structures before being processed by the application.

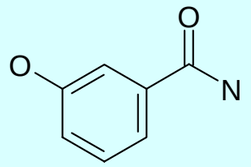
When indexing fragments, the system generates automatically XML index files, enriched with the information extracted from the MOL files and the name codes. Different levels and kinds of chemical information are then ordered and structured, ready to be extracted and analysed. When UserDB molecules have been transformed (using the vectors-descriptors) the sim/div computation can be achieved.

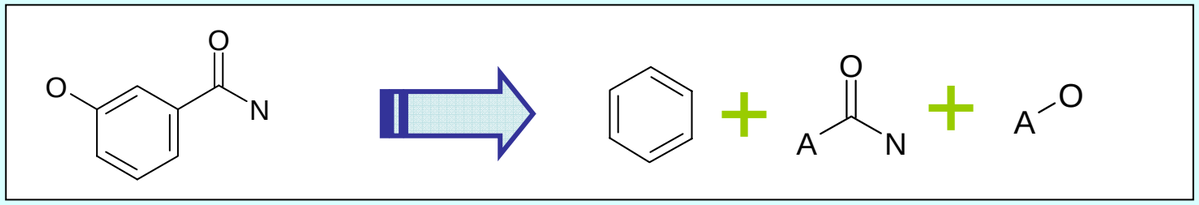
MOLECULAR SIMILARITY MEASURE

COMPARISON PROCESS

Once XML index file is filled and structured, **vector-descriptors** are constructed for each molecule by comparison of the hashed molecule with the FragDB. These representations are then compared following a defined algorithm and using different **indices** (Tanimoto, cosine, etc). Different levels of comparison should be possible: structure, structure + property and structure + property + position of the groups.

Test molecule: 3-Hydroxy-benzamide
Molecular weight : 137,14
Formule : $C_7H_7NO_2$



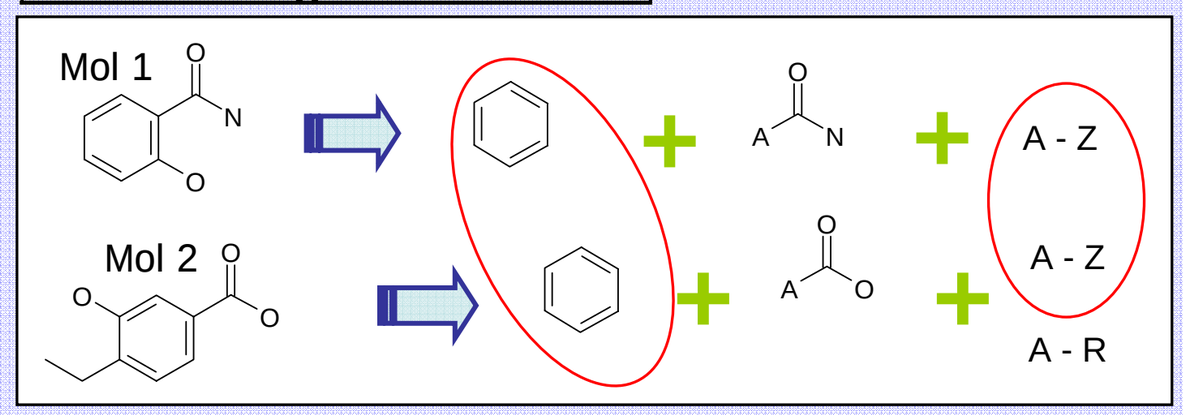


< 3-Hydroxy-benzamide;<CAUC6-54, AGCM-29c, AGSO-24 >>

Vector generation for a test molecule using FragDB substructure information

Index	Equation
1) Tanimoto	**St=c/(a+b-c)**
2) Simpson	Ss=c/min(a,b)
3) Cosine	$Sc=c/(a*b)^{1/2}$

Where the variables are:
a = fragments of the molecule 1
b = fragments of the molecule 2
c = common fragments between molecule 1 and 2



a	b	c
3	4	2

St=c/(a+b-c)
St=2/(3+4-2)
St=0,4

Simplified example of the Tanimoto Index computation for a couple of molecules

The Sim/Div measures then computed can be interpreted by the user to find new molecules having common substructures or properties with a defined query, to optimise a reactants database, to make a group of molecules more heterogeneous, to design a new test database, as well as, further application in Virtual Screening.

CONCLUSIONS

- Development of a system for the analysis of molecular diversity and similarity (MolDIA).
- Possibility to custom the analysis in different levels of comparison and different measures of diversity/similarity.
- Implementation of a notion of diversity that includes structural information and physicochemical properties in form of fragmentary vectors.
- Use of markup languages as a convenient way to represent and structure chemical information.
- Implementation of XML files to optimise the integration, query and management of diverse molecular data.
- Wide application fields: virtual screening, library design, database optimisation, etc.

REFERENCES

- Ana G. Maldonado, Bo Tao Fan and Michel Petitjean, "Using XML for structuring the chemical information: Towards a chemical knowledge representation". Proceedings of FIS2005, Third Conference on the Foundations of Information Science, Paris, July 4-7, 2005. http://www.mdpi.org/fis2005/proceedings.html.
- Ana G. Maldonado, Michel Petitjean, Jean-Pierre Doucet, Bo Tao Fan, "Molecular Similarity and Diversity: Concepts and Applications". Review, *Mol. Diversity*, 2005, in press.
- Jianhua Yao, Botao Fan, Jean-Pierre Doucet, Annick Panaye, Shengang Yuan, and Jianfeng Li, "SIRS-SS: A System for Simulating IR/Raman Spectra. 1. Substructure/Subspectrum Correlation", *J. Chem. Inf. Comput. Sci.* 2001, 41, 1046-1052.

Contact: Prof. B.T. Fan, fan@paris7.jussieu.fr
Dr. A.G. Maldonado, ana.maldonado@laposte.net