MolDiA: an XML-based System of Molecular Diversity Analysis Towards a Substructure Approach

Nowadays, in modern molecular discovery systems, a large number of compounds are tested and optimized in order to identify a molecule or group of molecules that act favorably in some desired circumstances. Sometimes, the key limiting factor in identifying new candidates is the availability of diverse collections of chemical compounds. When collections of chemical compounds are available, they tend to be enough large and unstructured (file format, chemical information, molecular properties, etc) to limit considerably the “human” treatment of these data.

From this, it is clear that the organization and the extraction of information and knowledge from molecular datasets is a key for the research of novel molecules. Techniques like virtual screening or data mining have been proposed recently to fulfill these needs. Similarity analysis engines are often used to compute resemblances between molecules in order to analyze and organize these databases, because researchers understood the potential interest to look for *similar* molecules rather than for *new* compounds. One way of finding similar molecules is to look at the resemblances of their structures.

The dissimilarity between two chemical structures can be measured as the differences between molecular graphs, and the *Similarity Property Principle* plays an important role in this task. Actually, if structurally similar molecules are more likely to have resembling properties, structural decomposition of the molecules seems to be a reliable basis for construction of molecular descriptors and computation of physicochemical properties.

Within this framework, the main objective during my PhD thesis consisted on the design and the implementation of a new chemoinformatic system based on a novel concept of molecular diversity. That’s how the MolDiA (Molecular Diversity Analysis) system was born.

MolDiA aims to the calculation and analysis of similarity and diversity of chemical databases in a chemo-structural framework. This analysis is done by means of weighted structural-property molecular descriptors which will allow us to compare the molecular datasets.

To build the molecular descriptors, it is necessary to decompose the molecules belonging to the datasets being analyzed. This decomposition follows defined rules (Figure 1) and allows us to compare the list of obtained fragments with a group of selected substructures belonging to a particular base, the FragDB.

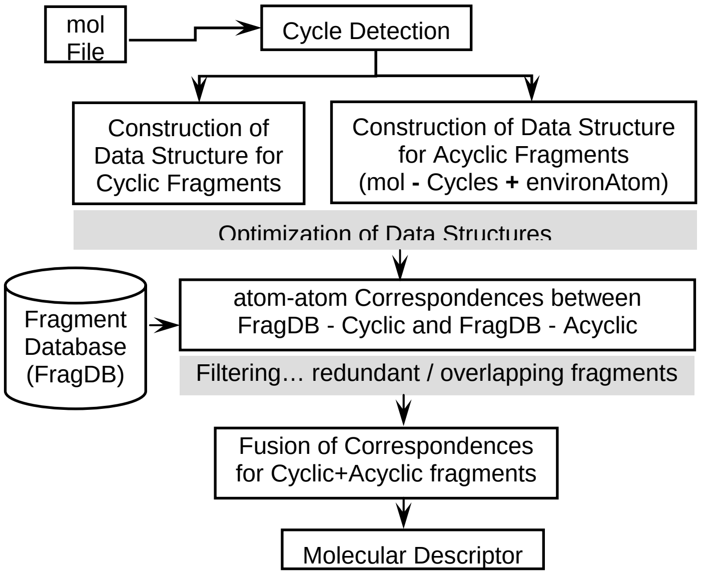

mol File
Cycle Detection
Construction of Data Structure for Cyclic Fragments
Construction of Data Structure for Acyclic Fragments (mol - Cycles + environAtom)
Optimization of Data Structures
Fragment Database (FragDB)
atom-atom Correspondences between FragDB - Cyclic and FragDB - Acyclic
Filtering… redundant / overlapping fragments
Fusion of Correspondences for Cyclic+Acyclic fragments
Molecular Descriptor

Figure 1. Generation of the molecular descriptor

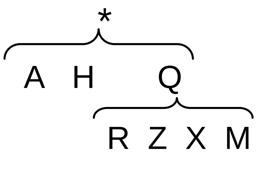
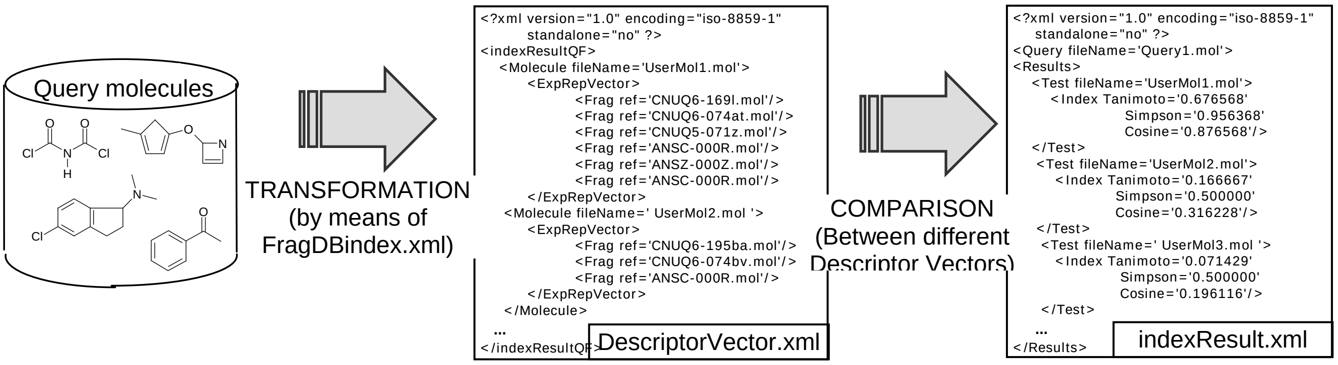


Figure 2. Generic atom hierarchy. The symbol * represents *all* the atoms. This category can be divided in aromatic atoms (A), Hydrogen atoms (H) and non aromatic, non H atoms (Q). The generic atom type Q can be classified in metals atoms (M), halogens (X), important heteroatoms (N, O, S, P, B) and the R group, which contains among others the Carbon and Silicium atoms.

The FragDB consists in a group of fragments predefined manually with the objective of covering the largest molecular space. In order to accomplish this goal, a list of common functional groups has been enriched with statistical information of frequent fragments, structures, atoms, etc. Then, we proceed to the banalization of linking atoms and some heteroatoms by the use of a hierarchy of generic atoms (Figure 2). Once the substructures composing FragDB have been chosen and properly defined, we index these files in the database using an XML file (see the first part of Figure 3). This file is generated automatically based on information extracted from molecular files, filenames and predefined properties rules.

To improve the indexation process, we have defined an encoded file name, which contains chemical and structural information, search keys, as well as, other information difficult to structure or extract later such as mixtures of heteroatoms or functional group presence-absence. The information encoded in the fragments filenames allows us to complement the data structure, improve the query of the database and in consequence the molecular analysis. Once the molecular descriptors have been constructed and indexed in another XML file, we proceed to their comparison in order to compute the molecular similarity and diversity (see the second part of Figure 3).



```
<?xml version="1.0" encoding="iso-8859-1"
        standalone="no" ?>
<indexResultQF>
   <Molecule fileName='UserMol1.mol'>
        <ExpRepVector>
                <Frag ref='CNUQ6-169l.mol'/>
                <Frag ref='CNUQ6-074at.mol'/>
                <Frag ref='CNUQ5-071z.mol'/>
                <Frag ref='ANSC-000R.mol'/>
                <Frag ref='ANSZ-000Z.mol'/>
                <Frag ref='ANSC-000R.mol'/>
        </ExpRepVector>
    <Molecule fileName=' UserMol2.mol '>
        <ExpRepVector>
                <Frag ref='CNUQ6-195ba.mol'/>
                <Frag ref='CNUQ6-074bv.mol'/>
                <Frag ref='ANSC-000R.mol'/>
        </ExpRepVector>
    </Molecule>
   ...
</indexResultQF>
```

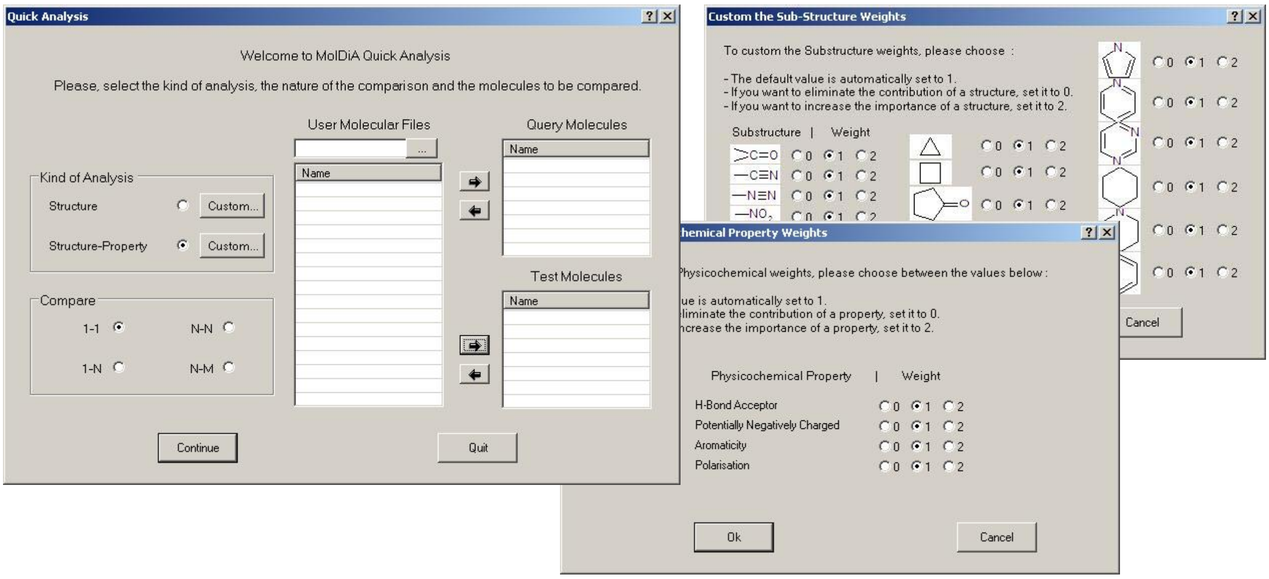
```
<?xml version="1.0" encoding="iso-8859-1"
    standalone="no" ?>
<Query fileName='Query1.mol'>
<Results>
   <Test fileName='UserMol1.mol'>
      <Index Tanimoto='0.676568'
                Simpson='0.956368'
                Cosine='0.876568'/>
   </Test>
    <Test fileName='UserMol2.mol'>
       <Index Tanimoto='0.166667'
               Simpson='0.500000'
               Cosine='0.316228'/>
    </Test>
     <Test fileName=' UserMol3.mol '>
       <Index Tanimoto='0.071429'
                Simpson='0.500000'
                Cosine='0.196116'/>
     </Test>
    ...
</Results>
```

Figure 3. Creation of an XML index of measures of similarity from a dataset of molecules, using the XML index of FragDB and the XML file of descriptors generated.

A descriptor comparison engine has been constructed, following the principle that strict/exact comparison between elements composing the descriptors can be complemented with a fuzzy comparison of the same elements. In this way, if a particular descriptor fragment is not present in the FragDB, a similar one can contribute to the similarity or diversity measure between the two studied molecules. Clustering of the substructures composing the FragDB was made to establish families of fragments, in order to implement the different levels of exact and fuzzy comparison of fragments when analyzing molecules. Another important feature of the system is the possibility to use structural and property weights in the calculus.

When computing the similarity between two molecules, it is possible to be more interested in a particular structure or functional group than other. Customizable weights for selected substructures belonging to FragDB are included in the MolDiA interface (Figure 4) to allow the chemists to choose by themselves the relative importance of particular fragments when analyzing a database. In the same way, some properties weights are attached to predefined fragments and are also customizable. The possibility to customize or modify the weighting scheme opens news insights in the treatment of different molecular systems using the same framework.

Different analysis and test made have shown the good performance of this new system; especially in drug-like databases (see Figures 5 and 6). “Small molecules” databases give less good results, in part because of the small size of the descriptors constructed. Different kinds of comparison are possible, using three indices of similarity: Simpson, Cosine and Tanimoto (possible data fusion to optimize the results). The introduction of molecular properties to the descriptor structure, as well as a whole system of weights, allow the tool to expand their use, from bibliographical structural research, to biomolecular target through generic molecules queries. The use of XML has proven to improve the data management of the system, the web interactivity and has opened a window for future implementations in chemistry and in science. Although the application of modern molecular diversity seems to be restricted to the identifications of new pharmaceutical molecules, the MolDiA approach can be applied to any endeavor that will likely benefit from the availability of increased molecular diversity: material science, flavors and fragrances, agrochemicals, catalysis, etc.



Quick Analysis

Welcome to MolDiA Quick Analysis

Please, select the kind of analysis, the nature of the comparison and the molecules to be compared.

Kind of Analysis
- Structure ○ Custom...
- Structure-Property ● Custom...

Compare
- 1-1 ●
- N-N ○
- 1-N ○
- N-M ○

User Molecular Files

Name

Query Molecules

Name

Test Molecules

Name

Continue | Quit

Custom the Sub-Structure Weights

To custom the Substructure weights, please choose :

- The default value is automatically set to 1.
- If you want to eliminate the contribution of a structure, set it to 0.
- If you want to increase the importance of a structure, set it to 2.

Substructure	Weight
>C=O	○ 0 ● 1 ○ 2
—C≡N	○ 0 ● 1 ○ 2
—N≡N	○ 0 ● 1 ○ 2
—NO₂	○ 0 ● 1 ○ 2

Cancel

hemical Property Weights

hysicochemical weights, please choose between the values below :

ue is automatically set to 1.
liminate the contribution of a property, set it to 0.
ncrease the importance of a property, set it to 2.

Physicochemical Property	Weight
H-Bond Acceptor	○ 0 ● 1 ○ 2
Potentially Negatively Charged	○ 0 ● 1 ○ 2
Aromaticity	○ 0 ● 1 ○ 2
Polarisation	○ 0 ● 1 ○ 2

Ok | Cancel

Figure 4. Different screenshots of the MolDiA Interface

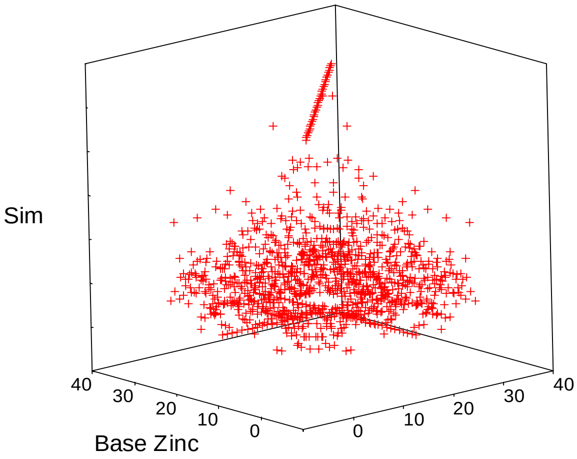


Figure 5. Diversity analysis of a dataset of 36 molecules from ZINC database. http://www.blaster.docking.org/zinc/

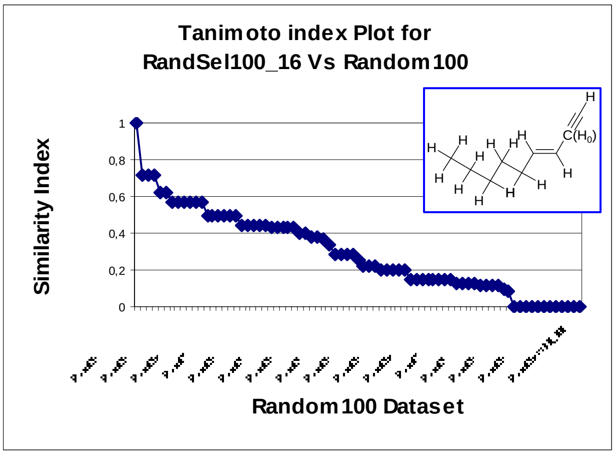


Figure 6. Similarity measures for a “small molecules” dataset of 77 structures *versus* a query molecule.

The originality of this approach is mainly due to the implementation of user-customizable weights for the structural and property information contained in the descriptors, as well as, an exact/fuzzy descriptor comparison engine, which can be enriched with new rules or constraints to improve the measures of molecular similarity and diversity. The inclusion of markup languages in order to manage the chemical information of the system is other novel feature which has improve the retrieve, extension, reuse and export/import of the different kinds of data necessary to the application.

Dissemination of the work achieved during the PhD thesis has been done. A state of the art on molecular similarity and diversity in chemoinformatics has been published as a comprehensive review [1]. Several results obtained with MolDiA during the last part of the PhD thesis have been published [2] or presented in conferences [3-6]. Finally, an up-to-date version of the PhD thesis book, defended last September, is online and available to the scientific community [7].

Currently, novel results and applications are being studied. Among others, we can mention the use of diversity matrices obtained with MolDiA to effectuate SVM analysis and clustering of molecular databases. Preliminary results show that in function of the level of the analysis granularity we can classify molecules following different criteria. Other applications are based on the substructure infrastructure developed for MolDiA, which can be used to compute similarity measures from customized graph comparison using kernel functions. Another study in project is the treatment of small molecular databases with measured activity data, in order to compare the ranking of MolDiA structural similarity obtained for a particular active query with an activity data given. This information can validate or optimize the parameters weights of the system for a particular problem, and serve as a model to effectuate drug/non drug analysis or clustering of bigger databases.

In parallel of these applications, two more publications are in preparation. The first explains the architecture and development of the system and the second give more insights about different results obtained using diverse datasets and the system of weights.

In order to achieve the applications which seem promising for the system developed during my PhD, I effectuate regular travels from Amsterdam to the ITODYS Laboratory in Paris in order to collaborate which the local research group, revise submitted papers and discuss about new implementation possibilities for the MolDiA system and possibilities of carrier in the national and international sphere.

References:

[1] Ana Maldonado et al. *Molecular Similarity and Diversity: Concepts and Applications*. Molecular Diversity, 10, 39-79 (2006).

[2] Ana Maldonado et al. *MolDIA: XML based system of molecular diversity analysis towards virtual screening and QSPR*. SAR and QSAR in Environmental Research 17, 11-23 (2006).

[3] Ana Maldonado et al. *A New Structural approach for Molecular Similarity and Diversity in Chemoinformatics*. Workshop Chemoinformatics in Europe: Research and Teaching. 29 May - 1 June, 2006. Obernai, France.

[4] Ana Maldonado *et al. MolDIA: Fragment Approaching to Molecular Similarity and Diversity Analysis.* The 12th International Workshop of QSAR in Environmental Toxicology. 8-12 May, 2006. Lyon, France.

[5] Ana Maldonado *et al. MolDIA: XML Based System of Molecular Diversity Analysis Towards Virtual Screening and QSPR.* The 3rd International Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2005). 29 Oct-1 Nov, 2005. Shanghai, China.

[6] Ana Maldonado. *Using XML for Structuring the Chemical Information: Towards a Chemical Knowledge Representation.* The 3rd Conference on the Foundations of Information Science (FIS2005). ENSTA. 4-7 July, 2005. Paris, France. Long paper published by MDPI. Online Edition ISBN 3-906980-17-0 (2005). http://www.mdpi.org/fis2005/proceedings.html

[7] Ana Maldonado. *Diversité Moléculaire : Application au Criblage Virtuel, Corrélation avec des Propriétés Physico-chimiques*. PhD thesis. University Paris 7 - Denis Diderot, 19 september 2006. Available (french) at: http://ana.maldonado.free.fr/perso/PhD_AnaMaldonado.pdf

Other references are available in the publication list or at the link: http://ana.maldonado.free.fr