MolDiA: A Novel Molecular Diversity Analysis Tool. 1. Principles and Architecture

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

ITODYS, Université Paris 7 − Denis Diderot, CNRS UMR-7086, 1 rue Guy de la Brosse, 75005 Paris, France

Received April 5, 2007

We introduce the principles and the architecture of a user-friendly software named MOLDIA (Molecular Diversity Analysis) which aims to the comparison of diverse molecular data sets through an XML structured database of predefined fragments. The MOLDIA descriptors are composed of complex fingerprint-like structures, which enclose not only structural information but also physicochemical property data. The system architecture includes the use of customizable weights on molecular descriptors and different choices of similarity/diversity measures to analyze the given data sets. Intermolecular comparisons using Ullmann's algorithm were optimized by the use of fuzzy logic, generic atoms, and a whole system of chemical graph analysis. We have found that customizing the similarity/diversity computation using structural and/or properties weights and choosing the level of fuzziness of the molecular comparison allow the user to adapt the tool to particular needs and increases the possibilities of MolDiA applications. The implementation of XML Web technologies has proven to improve and ease the extraction, processing, and analysis of chemical information.

INTRODUCTION

Let us imagine the situation where a medicinal, industrial, material, or other chemical researcher is looking for new molecules to enrich his/her collection of potentially interesting chemical compounds. Focusing on specific criteria (biomolecular targets, structural constraints, physicochemical properties, activity...), the researcher can have an idea of the desired properties, characteristics, or features of the molecules as well as the undesired ones. If commercial, academic, or industrial molecular databases are available for querying (sometimes the key limiting factor in identifying new candidates is the availability of diverse collections of chemical compounds), they tend to be unstructured and of large size. In consequence, human selection and test, in an attempt to identify a lead compound that acts favorably in some desired circumstances, can be a hard task. Later, once the leads have been identified, they rarely possess the whole set of properties, characteristics, or features that makes them useful. So they must be systematically optimized by varying their functional groups in order to decrease potential toxicity and improve potency, selectivity, oral bioavailability, and other criteria important for the researcher. Nevertheless, disposing of an efficient tool to help in the quest of new leads is extremely valuable to save time and money.

These situations have given new insights for further research areas involved in information management and automatic analysis, retrieval, and management of huge amounts of data, such as high throughput screening (HTS) and data mining in chemoinformatics. The organization and the extraction of information and knowledge from molecular data sets are a powerful key on the research of novel molecules.

Similarity and Diversity Analysis Using 2D Descriptors. 'Similarity analysis motors' are one of the approaches used to compute resemblances between molecules, in order to organize and query big databases. Researchers understood the potential interest to look for *similar* molecules rather than *new* compounds. Similarity analysis provides a simple and popular method for virtual screening and underlies the use of clustering methods on chemical databases. Furthermore, diversity analysis explores the way of how groups of molecules cover a determinate structural space and underlies many approaches for compound selection and design of combinatorial libraries. In similarity searching, a common approach is to use a 2D description of molecules by means of graphs or molecular environments. This approach has the advantage to be less computer time demanding (in comparison with 3D approaches) since the atom−atom comparisons algorithms are simpler. The conformational problems observed with higher dimensional descriptions and the stereoisomer comparison constraints (due to spatial geometric information) are avoided as well. Recent contributions[1−4] affirm that 2D substructure descriptors are able to establish relationships between the molecular information and given biological properties.

Chemical graphs are popular 2D representations used in different ways to analyze the chemical systems. They can be used by strict comparison (graph homomorphism) or partial comparison (graph isomorphism) of molecules, by molecular decomposition of substructures (to compare, analyze, recompose molecules), and by comprehension of structural paths (environmental approach).

The *substructure searching*[5,6] (or graph isomorphism) allows the screening of chemical databases by means of different structural criteria. In the *fragmental method*[7−10,19] molecules are decomposed in essential parts following predefined rules, to improve the design of new compounds and for the prediction of physical and biological properties through QSPR, QSAR and SAR models. The *reduced graph*[5,11,12] techniques transform the structural chemical information into a simplified representation for easing the

* Corresponding author phone: +33 1 44274857; fax: +33 1 44276814; e-mail: petitjean@itodys.jussieu.fr.

† This paper is dedicated to the memory of Prof. Bo Tao Fan.

10.1021/ci700120v CCC: $37.00
Published on Web 11/03/2007

virtual screening (for example, a phenyl ring can be reduced to its center represented by a "Ph" symbol).[5,11,12] A broad range of studies has been done[4] using graph representation and substructural approaches to address the virtual screening of chemical databases.

Another technique which has become a common way used by chemists in the similarity and diversity searching is the *molecular environment approach*.[5,11] It represents a molecule as a function of its ordered atomic/fragment environments. Ordered substructures are determined concentrically around a focus: atom, bond, substructure, etc. Among the known molecular environment approaches, some propose the customization of the focus environment according to the desired "depth" of the analysis and the description precision.[13,14] This approach offers the advantage to choose the focus depending upon the studied property as well as the depth of the environment which can be generated algorithmically in an optimized way. Other approaches are essentially atom centered[15] (neither bond nor structure centered). Sometimes, they exploit molecular environment information layer by layer (starting from the central atom of the target) to build a structural code.[16−18] In other situations the *atom environments* of molecules are used together with descriptors for similarity searching.[15] These descriptors are of easy interpretation and are very similar to the *signature molecular descriptors*.[20,21] They are calculated from the connectivity table which is constructed with an information gain based feature selection. The molecular atom environment fingerprints are binary indicators of count vectors of atom types. The distances (or layers) are computed from the central atom. These methods have in common the objective of describing 2D structures as an ensemble of complementary substructures, in order to ease the library management and analysis.

Apparition of tools having in common 2D molecular representation and Web compatibility is increasing rapidly. Most of these software are small and standalone programs, easy portable, and highly compatible with the Web (e.g., online databases and queries). Their "limits" in treating only 1D-2D information are highly compensated by user-customizable approaches which include most of the time chemical properties and features commonly used to parameter, filter, and personalize the search. MolDiA enter in this first category. Other examples are MolCart (MolSoft), PubChem (NCBI), ChemFinder (Cambridge Soft) Chem WebBook (NIST), JChemBase (Chemaxon), etc. Exceptionally standalone Web tools using 2D-3D representations are HyperChem DB (HyperCube), PDB (RCSB), and NCI DIS 3D Chem-X (Chem Design).

It is important to point out that not only standalone Web tools have used with success 2D-3D representation of molecules in similarity or diversity computations. However, for most of the commercial or free software "packages" the similarity module is a small part of a bigger framework. This second category of software has as main characteristics a high computing capacity, local chemical databases, and robust 3D representations and manipulations. Examples of these softwares are as follows: C^2 Diversity Module (Accelerys), MOE Sim Module (Chem Comp Group), SYBYL Surflex-Sim Module (Tripos), Glide Sim Module (Schrodinger), MDL QSAR Sim Module (MDL), EON & ROCS Sim Modules (OpenEye), etc.

In this paper, we introduce MolDiA, a novel chemoinformatic tool which aims to the calculation and analysis of molecular similarity and diversity in a 2D chemo-structural framework. In the first part of the paper, we explain the MolDiA general strategy for computing the similarity and the diversity measures of molecules. We will detail the construction and management of the system databases, the building and comparison of structural descriptors, and finally the similarity/diversity computation formalism. Then we illustrate the customization possibilities of the tool with a practical example. In the second part of the paper we discuss the choice of the approaches implemented, some preliminary results, and we explore the ways to exploit the full potentiality of the tool. Finally, we give some concluding remarks and perspectives.

METHODOLOGY

The general architecture of the MolDiA software is described in Figure 1. The system is mainly composed of three parts: an optimized fragment database, indexed and structured using XML, a comparison mechanism using a modified Ullman's algorithm to construct structural descriptors, and a descriptor comparison motor, which allows computing the molecular (dis)similarity of a query molecule respect to a reference database.

Fragment and Molecular Data Sets. MolDIA databases have been created with different goals and are structured at different levels. The first group of databases is called *user-managed databases* and is composed by the group of molecules to be screened (QueryDB) and a comparison database (CompDB). Once the user has selected the composition of both QueryDB and CompDB, the similarity/diversity analysis is made using a predefined *substructure database* named FragDB, the second and main group of molecules that compose the tool. The molecules of QueryDB and CompDB will then be described in terms of the fragments defined in FragDB. This database has been built using a hierarchy of generic atoms and a fragment code, as is described below.

Generic Atom Hierarchy and Encoded Fragment Nomenclature. This hierarchy was constructed to take into account the widest chemical diversity when building the fragments of FragDB. The hierarchy allows fuzzy analysis of the molecular systems according to different structural criteria. Generic atoms have been used for all the connecting atoms of the fragments and for some atoms in the heterocyclic systems. Main categories are divided in different subcategories as shown in Figure 2. In this scheme the symbol (*) represents *all* the atoms, (A) aromatic atoms, (H) hydrogen, and (Q) the non H, nonaromatic atoms. The generic atom type Q can be classified in metals and others atoms (M), halogens (X), important heteroatom (Z = N, O, S, P, B), and the R group, which contains the C and the Si atoms.

Once the substructures composing FragDB have been chosen and properly defined using generic atoms, we build the molecular files and index them in the database. To improve the indexation process, we have defined an *encoded file name scheme* (see Figure 3). The code contains chemical and structural information difficult to structure or to extract later, such as aromaticity, cycles containing multiple Z heteroatoms, membership to clusters of chemicals (family),

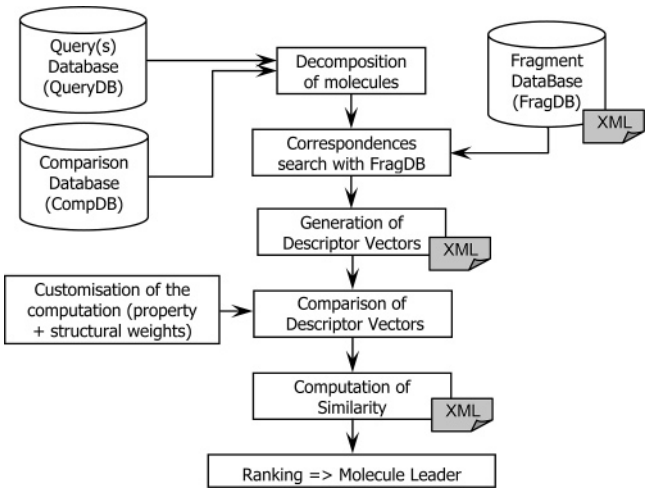

Figure 1. MolDiA general architecture.

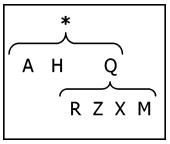

Figure 2. Generic atom hierarchy of MolDiA fragments.

etc. The information encoded in the fragments filenames allows the complementing of the data structure and improves the query of the database and in consequence the molecular analysis.

Construction and Clustering of the FragDB. The *similarity property principle*[22,23] is one of the basic statements used when constructing the FragDB. This principle states that structurally similar molecules tend to behave in a similar way. If molecules sharing the same group of fragments or substructures (and more important, the same functional groups) are expected to have similar properties, structural decomposition of the molecules seems to be a reliable basis for the construction of descriptors[23] and the computation of physicochemical properties.[19] To construct the building blocks of the FragDB, we have followed the next steps.

First, we have searched for published contributions[24−27] that could provide lists of common fragments, structures, atoms, etc. extracted from different molecular databases. Interesting substructures and functional groups have been collected by hand in order to complete the main list with fragments of pharmaceutical or medicinal interest. Then, we proceed to the generalization of linking atoms and some atoms composing common heterocycles by the use of generic atoms. The final FragDB list consists of a *nonexhaustive* set of substructures susceptible to correspond with fragments extracted from the molecular decomposition analysis. These fragments constitute the future elements of the molecular description represented under the form of descriptors-vectors. The quality and accuracy of the molecular descriptors highly depend on its composition.

The size of the fragment database has been stabilized to reach the number of 563 fragments. The number of fragments has evolved continuously to include new structures and improve those existent. Even if the size of the MolDiA fragment database is small it covers a big percentage of the fragments belonging to druglike molecules being tested. It is possible because of the generic nature of some of the fragment atoms and the criteria chosen to build the FragDB (high frequency/common fragments and functional groups). The construction of a *complete* and *finite* set of the "most representative" fragments for the whole space of chemical diversity is currently an unsolved problem and will certainly be one of the more challenging tasks for computational chemists of tomorrow. In a recent contribution, Ertl[29] has automatically created a large library of nearly 600 000 heteroaromatic scaffolds using only 8 atom types and 14 simple squeletons, in order to identify biological activity in selected data sets. Even if the ring generation obeys chemical rules to avoid unstable or "exotic" structures and to improve synthetic feasibility, there is not a guarantee that the rings created exist or are possible to synthesize. Another finding of the publication is that despite the huge size of the ring systems generated, only a very limited number of aromatic scaffolds have been found in the bioactive molecular collection studied. In another publication, Degen[28] introduces the FlexNovo tool which use a fragment space of approximately 4500 fragments to effectuate structure-based molecular design. The use of well-defined connection rules, a deterministic algorithm, and various filter criteria produces a comprehensive and user-customizable virtual molecular tool.

Once the FragDB has been designed, different fragments have been regrouped in the form of clusters (fragment families). The criteria to define the families are mainly structural, but other criteria can be used. The clustering is done, in order to implement different levels of exact and fuzzy comparison of substructures when analyzing molecules. For example, a tertiary amine is in the same family that a secondary or primary amine is, even if they do not have the same number of heavy atoms. The levels of fuzziness can vary from one fragment to another one. Thus, from certain fragments only the exact comparison will be allowed (e.g., cyano), while for others structural fuzziness can cluster several molecules (e.g., single or multiple substituted n-cyclohexane). Eighteen *families* of fragments have been created until now.

Four types of fragments compose the FragDB. Up to now, the FragDB is composed of 58 cyclic aromatic fragments (CA), 444 cyclic nonaromatic fragments (CN), 11 fragments containing acyclic chains (AN), and 50 fragments containing functional groups (AG). Some examples of FragDB substructures are given in Figure 4.

Even if the FragDB is given "as it" when running the MolDiA software, it is possible to modify and extend it, in the future. The use of XML for building the architecture of the databases as well as the flexible nature of the filename nomenclature allows and encourages the addition of new fragments to the FragDB. However, this process should be supervised (manually or automatically) to avoid adding duplicate, redundant, or superimposed fragments. A big FragDB is not a guarantee of having a wide structural scope. But a wise selection of representative fragments, is a good start for analyzing generic structural differences.

It is important to note that automatic tools for substructure extraction have been published,[16,30−32] but a high degree of superimposed and redundant structures is commonly found when using systematic extraction tools, whereas the MolDiA extraction motor refers to predefined substructures, optimized filters, and algorithms, which avoid this problem.

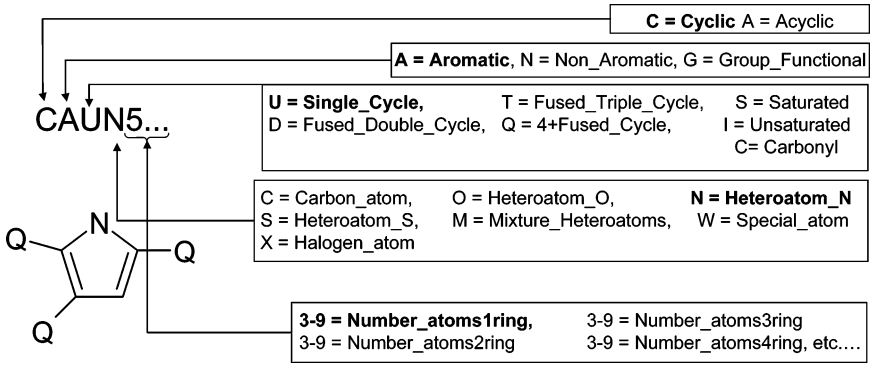

Figure 3. Fragment file name nomenclature

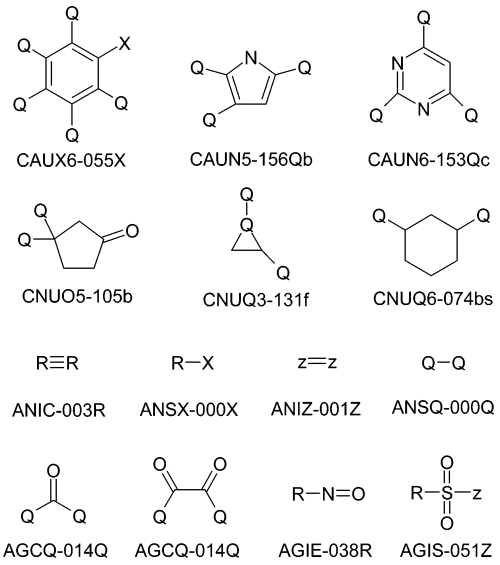

Figure 4. Examples of cyclic aromatic (CA), nonaromatic fragments (CN), acyclic fragments (AN), and functional groups (AG). Q, any heavy atom excluding aromatic atoms; R, carbon atom; X, halogen atoms; and Z, important heteroatoms (N, O, S, P, B).

MolDiA Descriptors: From Molecules to Vectors. The easiest way to do an accurate computation of the (dis)-similarity between two molecules is by their direct comparison. Unluckily this is not possible, since the concept and definition of "molecule" tends to be very diverse.[4] So, while the only accurate molecular description found until now (the exact resolution of the wave function equations for the molecular system) is not successfully solved on a manageable computer time frame, we have to deal with alternative, approximate, and simplified molecular representations. The MolDiA descriptors vectors share some characteristics of a *fingerprint*-type descriptor. Most of the software that addresses chemical diversity uses some forms of molecular fingerprint to describe each compound in a population.[33] Fragment based bit-string or fingerprints are widely used because of their computational efficiency and robustness. In MolDiA they are constructed from the chemical and structural information of the molecules to be analyzed, but differently to *classic* fingerprints, the number and nature of 'bits' on the fingerprint is not fixed or pre-established. The characteristics contained by the descriptor depend on the molecular information automatically decomposed in entities of a particular meaning (e.g., Si atom, functional group, aromatic ring, etc). In consequence, the size of the MolDiA descriptors is finite and limited to the size of the FragDB elements. Bigger molecules will have the tendency to produce larger descriptor-vectors than small molecules (which is one of the reasons—discussed in the last section—why the MolDiA framework works better with medium size molecules). To decompose the molecules, we fragment them in atomic units enriched with links to the environment information, as shown in Figure 5. This information is then compared with the fragment database, until all the atoms have been identified.

In order to build the descriptors-vectors, the analysis of the intramolecular information follows two steps. First, the smallest set of smallest rings of the molecule is identified using an in-house modification of the SSSR algorithm;[34] then the atoms belonging to the "boundaries" of the cyclic fragment are banalizated to increase the matching chances with the cycles of FragDB; and finally the optimized structures are compared using different levels of comparison (exact + fuzzy) with those of the FragDB. This comparison is done using an in-house optimized version of the Ullmann algorithm. The algorithm effectuates efficient graph isomorphism between two structures, by comparing the molecular characteristics of atoms and by using defined criteria to decide whether or not an atom can be equal or equivalent to another one. Sometimes duplicate fragments can be detected, if the atomic characteristics are identical but their position symmetric. In other cases, the fragments are just equivalents (certain atomic characteristics coincide according to definite rules) or superposed (when just one part of the substructures corresponds to the compared substructure). Modifications to the original algorithm have been done to accept a certain level of fuzziness in the atom comparison; in this way, different rules which allow comparing different classes of generic atoms between them and with real atoms have been added. Automatic detection of duplicate, equivalent, and superposed fragments as well as optimization and filtering algorithms are used at the final stage to minimize the number of detected fragments and to increment the number of correspondences with the fragments of FragDB.

The second step involves the analysis of the noncyclic parts of the molecule. The acyclic fragments are easily extracted from the original molecule by ignoring the atoms belonging to cycles. The connectivity of the acyclic fragments is rebuilt, and the "border" atoms (those who linked to the cyclic fragments) are banalizated to increase the matching chances with the acyclic structures of the FragDB. The optimized acyclic structures are then compared to the FragDB using an in-house modified version of the Ullmann algorithm and following different levels of exact and fuzzy

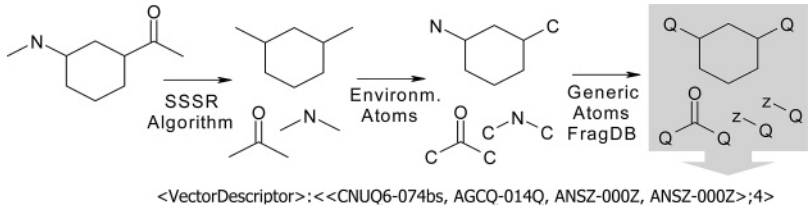

Figure 5. Molecular decomposition procedure: the molecule is first decomposed in cyclic and acyclic structures, then the close environment atoms are added, and finally the atoms are transformed in generic type atoms compatible with those of the FragDB.

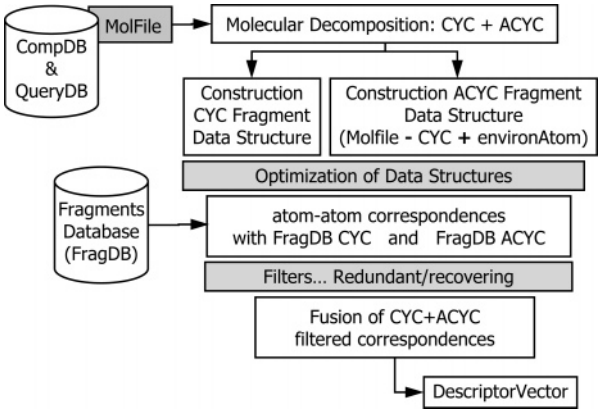

Figure 6. Construction of a descriptor-vector using cyclic and acyclic information.

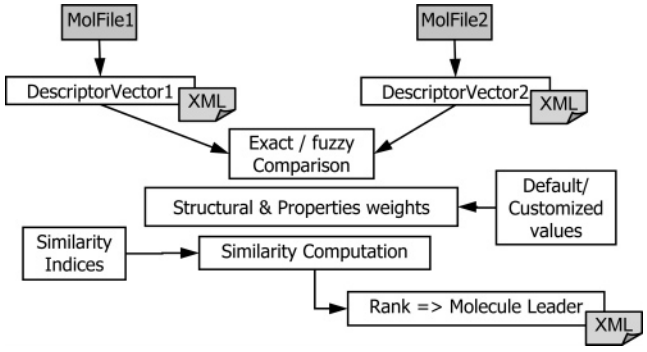

Figure 7. Descriptor comparison flowchart and computation of the similarity.

comparison as it was explained with the cyclic molecules. If the molecule to be analyzed is acyclic, then the first step is ignored, and the whole molecule is treated as a single acyclic fragment. Then, using atom−atom comparison with the acyclic fragments of the FragDB, matching structures are identified and extracted until no atoms last on the original molecule.

Once this last task is achieved, the matching fragments (cyclic and acyclic) are regrouped in a single structure to form the descriptor-vector representing the original molecule (see Figure 6). More details dealing with this procedure have been the subject of a recent report.[35]

Comparison of Descriptors: Computation of Similarity and Diversity. When comparing molecular representations (descriptors), it is necessary to quantify the degree of their chemical resemblance in an easy and reliable way. The similarity indices[36] allow us to implement functions that transform pairs of compatible molecular representations into real numbers, usually lying on the unit interval. Among the large number of similarity measures found in the literature,[35−38] a lot of them are interdependent or complementary. The use of multiple indices or coefficients when measuring the similarities/differences between two sets of descriptors lead to a better understanding of the relationships among the molecules.

In the last sections have been detailed the construction of the MolDiA descriptors using a predefined fragment database that applies a system of generic atoms, hierarchies, and clusters to improve the molecular description. Once the molecular descriptors have been constructed, their comparison quantifies the similarity and/or diversity between the molecules or groups of molecules. The descriptor comparison engine effectuates first a strict comparison between the elements belonging to the descriptor-vectors. Using the clusters (structural families) defined when constructing the FragDB it is possible to effectuate a fuzzy filename comparison between no-identical descriptor elements. The levels of fuzziness can vary from one fragment to another. If this comparison is not fully satisfactory (because they are elements that have not been strictly identified or do not belong to a family cluster), then a lower fuzzy level applies until all the fragments contained in the descriptor-vector are analyzed (if a particular descriptor fragment is not present in the FragDB, a similar one can always contribute to the similarity or diversity measure between the two studied group of molecules). Following this method, we have more chances to effectuate a reliable comparison of the structural descriptors.

Once vector entities are analyzed we use the information embedded within the descriptor-vectors to refine the measure of similarity and diversity using structural or/and physicochemical weights. The weights are attached to structural and property information embedded on the fragments corresponding to the molecular descriptor-vector. The discrete weight values (0, 1, or 2), have a particular meaning, and can be changed by the user through the software interface. The properties are extended (but not limited) to aromaticity, polarizability, hydrogen acceptor, and partial negative charge. More properties as well as more *family clusters* can be created and added at any moment by means of *rules* which can be automatically or manually added to the filters. The whole procedure is summarized in Figure 7.

The importance of introducing structural and properties weights in the framework resides in the expected scope of the similarity/diversity calculation. When computing differences or resemblances between molecules, it is sometimes useful to assign different degrees of importance to the various components of the molecular representations. For example under a particular context it could be important to find the molecules which possess an aromatic system or those which have two H-bond acceptor functional groups. Weights give the freedom to chemists to decide which substructure/

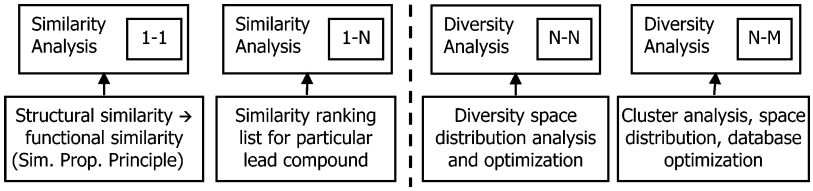

Figure 8. Similarity and diversity analysis proposed by MolDiA.

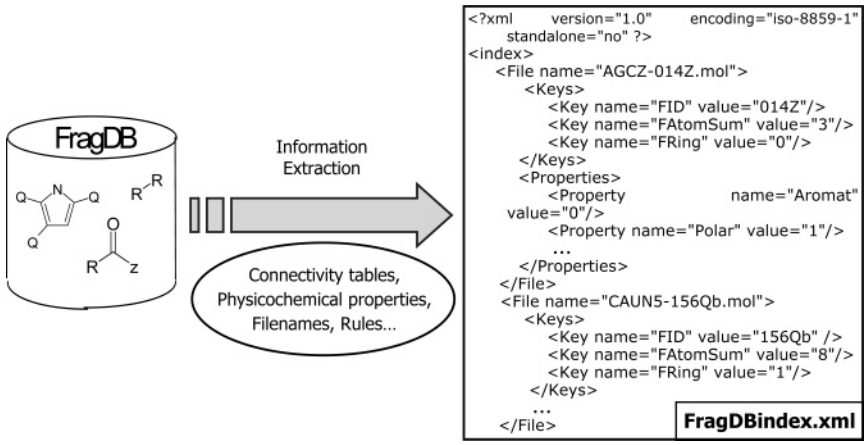

Figure 9. Creation of an XML index from the information extracted from FragDB.

property is more or less important when screening molecular databases.

User selected criteria establish important limits and clues to be observed when doing virtual screening. The possibilities to customize or modify the weighting scheme open new insights in the treatment of different molecular systems using the same framework. There are interesting publications about the effect of weighting schemes on the molecular similarity measures.[39,40]

Different kinds of analysis can be effectuated within the MolDiA formalism: *Similarity* analyses used to compare directly two structures (denoted as "1−1") or to establish a ranking list of one database in function of a molecular query (denoted as "1-N"). *Diversity* analyses used to identify molecules structurally different to another group of molecules or which have a particular fragment or property. If a group of molecules of a database is screened with itself (analysis denoted as "N−N"), then it is possible to analyze the distribution of the structures in the diversity space and eventually identify molecular clusters or diversity *holes*. If different databases are compared (analysis denote as "N−M"), then it is possible to study the composition, the number, and the nature of clusters present in one data set with respect to the other one. All these analyses are summarized in Figure 8.

XML for Structuring the Chemical Information. In the MolDiA system, XML files are generated at different stages of the molecular analysis. XML is the acronym for *eXtensible Markup Language* (XML),[41] a meta-language which proves to be a powerful tool for easing the information structure and exchange. Markup languages were, in principle, considered to be a universal format for structured documents on the Web. Today, they are widely used for representing any structured data and particularly scientific and hence chemistry data.

When working with the MolDiA databases we have encountered problems dealing with the accessibility, fast retrieval, and structuration of the data contained in the system. XML has proven to be an useful tool to solve MolDiA particular needs. The full details regarding the implementation of XML in MolDiA have been the subject of past publications;[42,43] in consequence, we only offer a schematic and brief explanation below.

First, the substructures contained in the FragDB are indexed using an XML structuring file. This file is generated automatically on the basis of the information extracted from molecular files, filenames, and predefined properties rules as shown in Figure 9. This information will be later used together with a parameter weight system to compute a user-customized similarity or diversity molecular measure.

The information enclosed into the XML file (e.g., key search tags and properties) can be extended or modified at any moment (by means of addition of new rules of extraction or calculation). These modifications have no negative incidence (slow down of the program execution, absence of broken link information) on the computation process of the similarity indices, since XML is a flexible and extensible language.

The index file is then parsed to build the descriptor-vectors which represent the query and the test molecules after the correspondences search with FragDB. The information contained in the index can be used at any moment to improve the matching process. Once each molecule has its corresponding descriptor, this information is stored in another XML file to ease the comparison process between vectors. The computation of (dis)similarity between two molecules is then done by comparison of their correspondent descriptor vectors. An XML file is automatically generated (Figure 10) to structure and organize the information correspondent to the analysis.

The index file is then parsed to construct the vectors descriptors which represent the query and the test molecules after the correspondences search with FragDB. The information contained in the index can be used at any moment to improve the matching process. Once each molecule has its

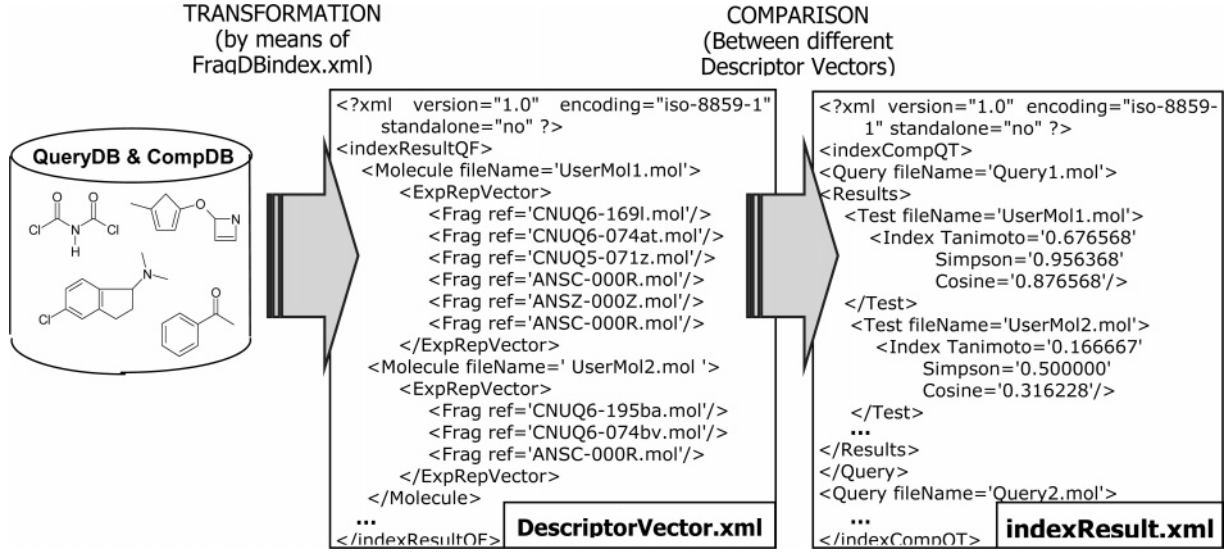

Figure 10. Creation of an index of results from QueryDB and CompDB molecules.

corresponding descriptor, this information is stored in another XML file to ease the comparison process between vectors. The computation of (dis)similarity between two molecules is then done by comparison of their correspondent descriptor vectors. An XML file is automatically generated (Figure 10) to structure and organize the information correspondent to the analysis.

The extracted and organized chemical information is the raw material for the molecular MolDiA (dis)similarity analysis. In the next section, the computation formalism of MolDiA is explained through some examples, and the effects of the weight implementation on the calculus of similarity and diversity measures are explained using some druglike molecules.

SIMILARITY/DIVERSITY COMPUTATION FORMALISM FOR VIRTUAL SCREENING

To explain the MolDiA theoretical formalism behind the general strategy for virtual screening we consider two molecules composed of $i = 1,...n$ and $i = 1,...n'$ fragments f_i, with the corresponding vectors of descriptors (or descriptor-vectors) V and V'.

$$V = (f_1,\ f_2,...,\ f_n)$$
$$V' = (f_1',\ f_2',...,\ f_n') \quad (1)$$

Considering that fragments belonging to the descriptor-vectors have *structural weights* attached, then each fragment intervening in eq 1 is multiplied by its corresponding relative weight.

$$V = (f_1w_1,\ f_2w_2,...,\ f_nw_n)$$
$$V' = (f_1'w_1',\ f_2'w_2',...,\ f_n'w_n') \quad (2)$$

When computing similarity between two molecules, we have to choose a similarity index. For example the well-known Tanimoto index with formula

$$S_T = \frac{c}{a+b-c} \quad (3)$$

where a represents the number of entities in the first molecule, b represents the number of entities in the second molecule, n is the total number of entities (dimension/size of the descriptor vector), c is the number of common entities, and d is the number of noncommon entities between the two molecules. An entity can be a structural, topological, or a physicochemical characteristic.

We can compute similarity and diversity measures taking into account user-customizable structural weights. Considering both *structural and property weights* (where *property* refers here not only to selected physicochemical properties but also to molecular features or characteristics attached to fragments) on descriptor-vectors, eq 1 transforms on a linear combination of different weights

$$V = ((\sum_j p_{1j}v_j) \times w_1,..., (\sum_j p_{nj}v_j) \times w_n)$$
$$V' = ((\sum_j p'_1j\ v_j') \times w_1',..., (\sum_j p'_{n'j}v\ j') \times w_{n'}') \quad (4)$$

where p_{ij} is the jth property or feature of the ith fragment of molecules V or V', v_j is the property weight, and w_i is the correspondent structural weight. For each fragment i, with j properties, we regroup the information concerning the physicochemical properties, the structural information and the weights, in an element called $\ll e_i \gg$.

$$e_i = w_i \frac{\sum_j p_{ij}V_j}{\sum_j} p_{ij} \quad (5)$$

Following the formalism, the $\ll a, b, c \gg$ values of the Tanimoto (or other similarity) index changes because not all of the fragments (regrouped in form on entities) contribute in the same way

$$a = \sum_{i=1}^{n} e_i,\ b = \sum_{j=1}^{n'} e_j,\quad c = \sum_{k=1}^{\min(n,n')} e_k \quad (6)$$

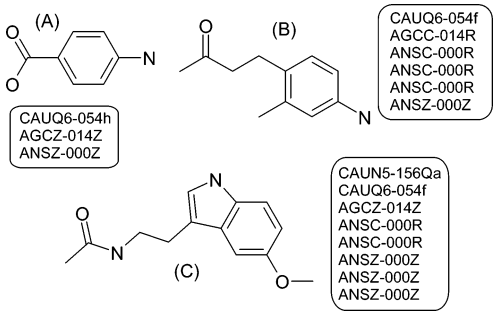

Figure 11. Three molecules and their correspondent MolDiA descriptor-vectors: (A) 4-aminobenzoic acid (PABA), (B) 4-(4-amino-2-methylphenyl)butan-2-one, and (C) N-[2-(5-methoxy-1H-indol-3-yl)ethyl]ethanamide (melatonin).

where e_k represents the common elements between e_i and e_j, and the sum is effectuated under all the elements composing the descriptor-vector for molecules V and V'.

The addition of isolated fragment contributions allows computing the global (dis)similarity index between the molecules V and V'. Multiples measures of (dis)similarity are also possible since S_T formula can be chosen. Virtual screening of chemical databases is then possible by implementing automatic treatment of molecular sets.

How Similar Are the Melatonin and the PABA Molecules? A Practical Example. To illustrate the formalism, let us consider the molecules of Figure 11. Following the molecular decomposition procedure described in Figure 5, the descriptor-vectors corresponding to these molecules have being computed. The relative weight (w_i) of each descriptor-vector element as well as the property/characteristic (p_{ij}) weights (υ_j) are customized, in order to study the relationships between the molecules of Figure 11 while using different criteria. For this example, two fragment features are considered: aromaticity (p_1) and H-bond acceptor (p_2). These features have been modified to explore the effect in the calculus of customized against default property-weights (Table 1a,b). Furthermore, we assign different structural weights to fragments of interest (Table 1c) in order to study the effect of these parameters on the similarity computation.

The first set of properties of Table 1a is called the "default set". It represents the default MolDiA output. We remark that the MolDiA software is not optimized for a particular group of molecules; however, the system was designed to allow the user to set his/her own parameters for specific needs. The default fragment contributions are equivalent and equal to one, for all the fragments of the analyzed molecules.

In Table 1b we suppose that "aromaticity" is an important criterion when comparing the molecules A, B, and C. In consequence, the weight corresponding to this feature has been set to two. We notice that the fragment contribution changes for the fragment concerned by the new requirements.

In Table 1c we modify again the criteria and consider multiple constraints: five-membered aromatic heterocyclic ring (pyrrole) and heteroatom-bonded carbonyl (as amide or ester) have more importance, whereas six-membered aromatic rings are ignored when comparing the molecules. Drastic changes are observed on the fragment contributions with respect to the default values.

Table 1. Customizable Weights and Molecular Properties for Some Fragments Belonging to Molecules A−C of Figure 11[a]

(a) default structural weight ($w_i = 1$), default property weights ($\upsilon_j = 1$)	w_i	p_1	υ_1	p_2	υ_2	$\sum_j p_{ij}\upsilon_j(w_i/(\sum_j p_{ij}))$
CAUQ6-054	1	1	1	0	1	1
CAUN5-156	1	1	1	0	1	1
AGCZ-014Z	1	0	1	1	1	1
AGCC-014R	1	0	1	1	1	1

(b) default structural weight ($w_i = 1$), customized property weight ($\upsilon_1 = 2$)	w_i	p_1	υ_1	p_2	υ_2	$\sum_j p_{ij}\upsilon_j(w_i/(\sum_j p_{ij}))$
CAUQ6-054	1	1	**2**	0	1	2
CAUN5-156	1	1	**2**	0	1	2
AGCZ-014Z	1	0	**2**	1	1	1
AGCC-014R	1	0	**2**	1	1	1

(c) customized structural weight ($w_i = 0, 1, 2$), default property weights ($\upsilon_j = 1$)	w_i	p_1	υ_1	p_2	υ_2	$\sum_j p_{ij}\upsilon_j(w_i)/(\sum_j p_{ij}))$
CAUQ6-054	**0**	1	1	0	1	0
CAUN5-156	**2**	1	1	0	1	2
AGCZ-014Z	**2**	0	1	1	1	2
AGCC-014R	**1**	0	1	1	1	1

[a] The first five columns give the fragment weights (w_i), the property values (p_1, p_2), and their corresponding weights (υ_1, υ_2); all of them are fixed by the user. The last column corresponds to the contribution of the fragment to the calculation of molecular similarity. Modified weight values are set in bold.

Adding the correspondent contribution for each element of the descriptor-vector (included or not in Table 1) and using eq 6, we obtain in Table 2 the similarity measures for each couple of molecules in function of modifications given to the weights.

Different similarity coefficients have been used to show the phenomena of index complementarity and to study their differences when applied to various molecular systems. Novel techniques use this information to compute "fusions" or combinations of different similarity measures for data ranking and data analysis.[44] The major differences related with the index used, move us to analyze in detail each index. The Simpson index is the least restrictive of the three used. The presence of common fragments is enough to identify the molecules as equivalent (fake isomorphism phenomena: unitary value for the A−C comparison). The Tanimoto index tends to give the more restrictive measures, since the presence−absence of noncommon fragments is one of the criterion used when comparing the molecules. The Cosine index values are intermediary. An extensive study made for different sets of molecules confirms these tendencies.[35,45]

Similarity computation using the default set of structural and feature weights (data Table 1a) shows the weakness of a nonparametric approach. Several problems arise when comparing the molecules A−C. The first and most evident is that since the descriptor-vectors have different sizes, even if they are normalized, the contribution of each fragment is not the same, except for particular cases. For small molecules (like the molecule A), the low number of fragments give more chances than big molecules (as molecule C) to find correspondences with other molecules. On the other hand, if big molecules are composed of a lot of fragments without functional groups (as small acyclic fragments, Q−Q, R−R, large unsatured carbon chains representing a lot of C−C fragments, etc.), then they will include a kind of "noise"

Table 2. Similarity/Diversity Comparison Indices for Molecules A−C of Figure 11, Using the Set of Weights of Table 1[a]

	weights Table 1a			weights Table 1b			weights Table 1c		
	T	C	S	T	C	S	T	C	S
A−B	0,29	0,47	0,67	0,38	0,57	0,75	0,14	0,26	0,33
A−C	0,38	0,61	1,00	0,44	0,63	1,00	0,33	0,58	1,00
B−C	0,40	0,58	0,67	0,42	0,60	0,71	0,27	0,45	0,60

[a] Tanimoto, Simpson, and Cosine indices[38] correspond to T, S, and C, respectively.

when considering the total number of fragments in eq 6. In these cases, the parametrization of these variables can, if not completely avoid, at least diminish the local effects. Structural correspondence values in Table 2 obtained using the data of Table 1a show the comparing nature of the MolDiA descriptors and the fact that the implemented algorithms have a tendency to be stricter than simple visual comparison. So, for molecules A and B, even if in appearance they share the same fragments, a detailed analysis reveals that connectivity, ordering, and nesting of the fragments is not equivalent. The same results are observed when molecules A−C and B−C are compared.

In order to correct problems due to a nonparametric approach, a system of property and structural weights has been included in the MolDiA system. When modifying the feature weights, one can think of customizing the similarity computation for a set of particular feature constraints (physicochemical properties, number and nature of aromatic systems, etc.). The number and kind of the selected properties or features can be modified easily on the system, thanks to the extensive nature of XML, as it has been pointed out before.

Modifying the aromatic feature weights (as shown in Table 1b) when comparing molecules A−C is equivalent to ignoring or increasing the comparison parameters. Like all the molecules studied have aromatic fragments, no significant differences can be observed on the similarity indices. An exception is the rise of the Tanimoto index for the A−C couple which identifies this couple as the more similar of the three with 44% (in contrast with the default values, where the B−C couple is the more similar with a rate of 40% in Table 2-1a). This can be due to the small size of A and the higher contribution of the aromatic ring. The higher similarity effect on couple A−C is clear when using the Cosine and Simpson indices of default (61% and 100%, Table 2-1a) and parametrized (63% and 100%, Table 2-1b) calculations.

To bypass the problem of high correspondence of small molecules (e.g., molecule A) and highly redundant fragments of big molecules (e.g., molecule C), the introduction of structure weights (Table 2-1c) increase the control on the comparison process. Assignment of a null structural weight (as for six-carbon aromatic rings in Table 1c) cancels its effect on the comparison. Molecules A−C are only being compared in function of the rest of the molecular skeleton. But like heteroatom-bonded carbonyl which has an increased effect on the comparison, molecules which *do not* have this element in common, suddenly lose similarity. For the couple A−B, 50% similarity loss is observed: from 29%; 47% and 67% (Table 2-1a) to 14%; 26% and 33% (Table 2-1c) for the Tanimoto, Cosine, and Simpson indices, respectively.

Arguments against similarity property principle using fragmental methods state that small changes in the chemical structure may lead to a dramatic change in the biochemical activity and/or in the ADMET properties.[46−49] In consequence, structural comparison of molecules could seem useless. But what is sometimes misunderstood is that biological activity is usually the result of the interplay of a number of complex processes, which cannot be easily represented by a set of linear relationships. To describe these processes better, nonlinear variable mapping can be used, where the activity is represented by a nonlinear function of structural, topological, and molecular descriptors.[9,50] In any case, there is considerable practical evidence[23,51−55] to suggest that the Similarity Property Principle can be considered as a rule and not an exception.

The definition of similarity itself is another point. It includes a number of variables that in function of their implementation can give affordable results or not, for a determined set of descriptors, properties, or molecules used in the study. Similarity results addressed in the bibliography are then dependent on the descriptor/property/similarity framework implemented.

CONCLUSIONS

The principles and the architecture of the program MolDiA have been presented and detailed. The general strategy consists mainly of a 2D structure-based approach which uses an advanced fragmental algorithm to decompose and analyze molecules. An original fragment nomenclature, the use of XML technologies, and a fuzzy atom comparison system allows constructing and structuring the software fragment database. All these methods, particularly the nomenclature research keys and the atom hierarchy, have proven to improve the quality and speed of molecular comparisons.

A novel feature in this molecular comparison system is the presence of a complex system of structural and property weights completely user-customizable. (Dis)similarity indices can be also personalized as a function of a chemist's particular needs. Undesirable functional groups or substructures can be ignored or certain particular fragments can be highlighted, when measuring differences between molecules. The real advantage of such a system is that it can be influenced and personalized by trained chemists to propose new molecular/weights arrangements and adapted to particular situations or data sets.

Although the application of virtual screening through molecular diversity seems to be confined to the identifications of new pharmaceutical molecules, this approach can be fruitfully applied to any endeavor that will likely benefit from the availability of increased molecular diversity: flavors and fragrances, agrochemicals, catalysis, material science, etc.

In a forthcoming paper,[45] we present the implementation of the MolDiA software which includes the development of a user-friendly interface, the integration of a flexible system of weights, and the construction of (dis)similarity matrix outputs. Tests run on MolDiA using a different set of molecules and weights and the performance evaluation of the tool in analyzing druglike molecules at different levels (1-N, N−N) will be presented.

REFERENCES AND NOTES

(1) Martin, Y. C. Molecular Diversity: How we Measure it? Has it Lived up to its Promise? *Il Farmaco* **2001**, *56*, 137−139.

(2) Gillet, V.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338−345.

(3) Bayada, D. M.; Hamersma, H.; Van Geerestein, V. J. Molecular Diversity and Representativity in Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.

(4) Maldonado, A.; Petitjean, M.; Doucet, J.-P.; Fan, B. T. Molecular Similarity and Diversity: Concepts and Applications. *Mol. Diversity* **2006**, *10*, 39−79.

(5) Barnard, J. M. Substructure Searching Methods: Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532−538.

(6) Luque-Ruiz, I.; Cerruelo-Garcia, G.; Gomez-Nieto, M. A. Representation of the Molecular Topology of Cyclical Structures by Means of Cycle Graphs. 2. Applications to Clustering of Chemical Databases. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1383−1393.

(7) Randic, M.; Wilkins, C. L. Graph Theoretical Ordering of Structures as a Basis for Systematic Searches for Regularities in Molecular Data. *J. Phys. Chem.* **1979**, *83*, 1525−1540.

(8) Randic, M. Graph Valence Shells as Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 627−630.

(9) Gorse, D.; Rees, A.; Kaczorek, M.; Lahana, R. Molecular Diversity and its Analysis. *Drug Discovery Today* **1999**, *4*, 257−264.

(10) Cuissart, B.; Touffet, F.; Cremilleux, B.; Bureau, R.; Rault, S. The Maximum Common Substructure as a Molecular Depiction in a Supervised Classification Context: Experiments in Quantitative Structure/Biodegradability Relationships. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1043−1052.

(11) Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic Identification of Molecular Similarity Using Reduced-Graph Representation of Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639−643.

(12) Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 13. Reduced Graph generation. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 260−270.

(13) Dubois, J. E.; Mercier, C.; Panaye, A. DARC topological system and computer aided design. *Acta Pharm. Jugosl.* **1986**, *36*, 135−169.

(14) Dubois, J. E.; Doucet, J. P.; Panaye, A.; Fan, B. T. DARC Site Topological Correlations: Ordered Structural Descriptors and Property Evaluation. In *Topological indices and related descriptors in QSAR and QSPR*; Devillers, J., Balaban, T., Eds.; Gordon and Breach Sciences Publishers: Amsterdam, 1999; pp 613−673.

(15) Bender, A.; Mussa, H. Y.; Glen, R. C. Molecular Similarity Searching Using Atoms Environments, Information-Based Feature Selection and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.

(16) Bremser, W. HOSE- A Novel Substructure Code. *Anal. Chem. Acta* **1978**, *103*, 355−365.

(17) Hull, R. D.; Singh, S. B.; Nachbar, R. B.; Sheridan, R. P.; Kearsley, S. K.; Fluder, E. M. Latent Semantic Structure Indexing (LaSSI) for Defining Chemical Similarity. *J. Med. Chem.* **2001**, *44*, 1177−1184.

(18) Xiao, Y.; Qiao, Y.; Zhang, J.; Lin, S.; Zhang, W. A Method for Substructure Search by Atom-centered Multilayer Code. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 701−704.

(19) Xing, L.; Glen, R. C. Novel Methods for the Prediction of Log P, pKa and Log D. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796−805.

(20) Faulon, J. L.; Visco, D. P., Jr.; Pophale, R. S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707−720.

(21) Faulon, J. L.; Churchwell, C. J.; Visco, D. P., Jr. The signature Molecular Descriptor. 2. Enumerating Molecules from their Extended Valence Sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721−734.

(22) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening - An Overview. *Drug Discovery Today* **1998**, *3*, 160−178.

(23) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(24) Stobaugh, R. E. Chemical Abstracts Service Chemical Registry System. 11. Substance-Related Statistics: Update and Additions. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 180−187.

(25) Xu, J.; Stevenson, J. Drug-like Index: A New Approach to Measure Drug like Compounds and their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.

(26) Sheridan, P. R. The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103−108.

(27) Ertl, P. Chemoinformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374−380.

(28) Degen, J.; Rarey, M. FlexNovo: Structure-based Searching in Large Fragment Spaces. *ChemMedChem* **2006**, *1*, 854−868.

(29) Ertl, P.; Jelfs, S.; Mühlbacher, J.; Schuffenhauer, A.; Selzer, P. Quest for the Rings. In Silico Exploration of Ring Universe To Identify Novel Bioactive Heteroaromatic Scaffolds. *J. Med. Chem.* **2006**, *49*, 4568−4573.

(30) Takahashi, Y. Automatic Extraction of Ring Substructures from a Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 167−170.

(31) Dubois, J. E.; Carabedian, M.; Ancian, B. Automatic Structural Elucidation by C-13 NMR - DARC-EPIOS Method - Description of Progressive Elucidation by Ordered Intersection of Substructures. *C. R. Hebd. Seances Acad. Sci. Ser. C* **1980**, *290*, 383−386.

(32) Carabedian, M.; Dagane, I.; Dubois, J. E. Elucidation by Progressive Intersection of Ordered Structures from Carbon-13 Nuclear Magnetic Resonance. *Anal. Chem.* **1988**, *60*, 2186−2192.

(33) MACCS keys, BCI fingerprints, MDL keys, and Unity Fingerprints information available at http://www.mesaac.com/Fingerprint.htm, http://www.bci.gb.com/prod_fingerprint.html, http://www.daylight.com/dayhtml/doc/theory/theory.finger.html, and http://www.tripos.com/, respectively (accessed Jul 8, 2007).

(34) Fan, B. T.; Panaye, A.; Barbu, A.; Doucet, J.-P. Ring perception. Application of Elimination Technique to the SSSR Search from a Connection Table. The first European conference on computational chemistry (E.C.C.C.1). AIP Conference Proceedings. April 5, 1995, Vol. 330, pp 538−543.

(35) Maldonado, A. G. Diversité Moléculaire: Application au Criblage Virtuel, Corrélation avec des Propriétés Physico-chimiques. Ph.D. Thesis. University Paris 7 - Denis Diderot, France, 2006. http://ana.maldonado.free.fr/index.html (accessed Jul 8, 2007).

(36) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, Herts, U.K., 1987.

(37) Maggiora, G. M; Shanmugasundaram, V. Molecular Similarity Measures. In *Methods in Molecular Biology*; Vol. 275, Chemoinformatics. Concepts, Methods and Tools for Drug Discovery, Bajorath, J., Ed.; Humana Press Inc.: Totowa, NJ, 2004; pp 1−50.

(38) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(39) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating Between Drugs and non Drugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(40) Bath, P. A.; Morris, C. A.; Willett, P. Effects of Standardization on Fragment-Based Measures of Structural Similarity. *J. Chemom.* **1993**, *7*, 543−550.

(41) *Extensible Markup Language (XML) 1.0*, 4th ed.; W3C Recommendation, August 16, 2006. http://www.w3.org/TR/REC-xml (accessed Jul 8, 2007).

(42) Maldonado, A.; Petitjean, M.; Doucet, J.-P.; Panaye, A.; Fan, B. T. MolDIA: XML Based System for Molecular Diversity Analysis Towards Virtual Screening and QSPR. *SAR QSAR Environ. Res.* **2006**, *17*, 11−23.

(43) Maldonado, A. *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 (accessed Jul 8, 2007).

(44) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.* **2006**, *46*, 462−470.

(45) Maldonado, A.; Doucet, J.-P.; Petitjean, M.; Fan, B. T. MolDiA: A Novel Molecular Diversity Analysis Tool. Part 2: Implementation and Performance. Manuscript in preparation.

(46) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity - a Review. *QSAR Comb. Sci.* **2003**, *22*, 1006−1026.

(47) Bajorath, J. *Virtual Screening in Drug Discovery: Methods, Expectations and Reality*. http://www.currentdrugdiscovery.com/pdf/2002/3/BAJORATH.pdf (accessed Jul 8, 2007).

(48) Turin, L.; Fumiko, Y. *Structure-Odor Relations: A Modern Perspective*. http://www.flexitral/research/review_final.pdf (accessed Jul 8, 2007).

(49) Meylan, W. M.; Howard, P. H.; Boethling, R. S.; Aronson, D.; Printup, H.; Gouchi, S. Improved Methods for Estimating Bioconcentration/Bioaccumulation Factor from Octanol/Water Partition Coefficient. *Environ. Toxicol. Chem.* **1999**, *18*, 664−672.

(50) Japertas, P.; Didziapetris, R.; Petrauskas, A. Fragmental Methods in the Design of New Compounds: Applications of the Advanced Algorithm Builder. *QSAR* **2002**, *21*, 23−37.

(51) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(52) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighbourhood Behaviour: a Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(53) Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407−1414.
(54) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.
(55) He, L.; Jurs, P. C. Assessing the Reliability of a QSAR Model's Predictions. *J. Mol. Graphics Modell.* **2005**, *23*, 503−523.

CI700120V