Determination of Screening Descriptors for Chemical Reaction Databases

Laurent Dury[*], Laurence Leherte, and Daniel P. Vercauteren

Laboratoire de Physico-Chimie Informatique,
Facultés Universitaires Notre-Dame de la Paix
Rue de Bruxelles, 61; B-5000 Namur (Belgium)
tel: +32-81-734534, fax: +32-81-724530
`firstname.lastname@fundp.ac.be`

Abstract. The development of chemical reaction databases has become crucially important for many chemical synthesis laboratories. However the size of these databases has dramatically increased, leading consequently to perfect more and more powerful search engines. In this sense, the speed and the efficiency of screening processes of the chemical information are essential criteria. Looking forward for powerful algorithms dedicated to information retrieval in chemical reaction databases, we have thus developed several new graph descriptors to find efficient indexation and classification criteria of chemical reactions.

1 Introduction

During the last two decades, organic chemists started to use computer programs in organic synthesis, especially "Computer Aided Organic Synthesis" (CAOS) programs which were developed to discover strategies for the preparation of target molecules. The use of such an approach, based on Corey's original concept of retrosynthetic analysis, was first demonstrated in 1969 by Corey and Wipke [1].

While keyword searching has been available in Chemical Abstracts (CAS) and elsewhere since the early 1970s, and (sub)structure searching in DARC and CAS ONLINE has been available since the beginning of the 1980s, "real" reaction databases, *i.e.*, containing structure diagrams, were implemented in the mid-1980s only (*cfr.*, for example, Reaction Access System (REACCS), MDL, 1982; Synthesis Library (SINLIB), 1982; Organic Reaction Access by Computer (ORAC), 1983).

Since 1990, a number of new reaction databases appeared such as CASREACT, ChemInform RX (CIRX), FIZ Chemie, ChemReact, and Beilstein CrossFire, and more importantly, the access to various reaction database servers was significantly improved. The current reaction databases incorporate all the elements and facilities of other types of chemical databases, such as text (for bibliographic data, reaction classifications) and/*or* numeric values.

[*] FRIA Ph.D. Fellow

D.A. Zighed, J. Komorowski, and J. Zytkow (Eds.): PKDD 2000, LNAI 1910, pp. 388−394, 2000.

As the size of all these databases, in terms of molecular structures as well as of reactions, is increasing drastically every year, the efficiency of searching algorithms implemented within these databases is thus primordial. Optimization of these algorithms usually involves the implementation of a searching process in two separate steps. First, the screening step which determines, with the help of a set of descriptors based on the user's query, a subset of potential candidates. In a second time, a procedure usually called ABAS (Atom by Atom Search) [2] compares all components of the subset with the user's query. The ABAS part is the most time consuming step. The speed of the search is indeed inversely proportional to the size of the candidate subset given by the initial screening; this step thus controls the efficiency of the search. Therefore, the choice of the screening descriptors is very important; the adequacy of that choice determines the performance of the database search.

There are two main families of chemical descriptors which usually depend on the user's request. First, if a reaction is searched on the basis of one its molecular component, a reactant or a product, the descriptor is expressed in terms of molecular structure or substructure features. These descriptors are called molecular descriptors. Conversely, if all constituants are known, the descriptor is based on the reaction pattern, *i.e.*, the smallest set of atoms and bonds which allows to describe the transformation occurring between the products and the reactants. These descriptors are called reaction descriptors.

In order to optimize a searching algorithm, for each component of the databases, all the descriptor values need to be calculated and stored in an annex database of pointers which refer to the searched reaction database.

In this short paper, we first describe some basics about the oriented-object programming use to represent and analyse the chemical reactions. Then, we present our ideas on the development of a powerful reaction searching algorithm by describing several molecular descriptors such as reduced graph representations of molecules and reaction descriptors based on the breaking and the formation of chemical bonds.

2 Oriented-Object Programmation

In our program, an oriented-object programming language is used to manipulate an organic reaction represented using three levels of objects. As the molecular structures involved in a reaction are constituted of atoms and bonds, the first level contains this information. Each atom and bond objects, named TAtom and TBond, respectively, are next grouped in a second level which describes the molecular structure and which is called TMol. Finally, the set of molecule objects associated with the reactant(s) and product(s) are grouped in a third level object symbolizing the organic reaction (TRxn). For example, as shown in Figure 1, related to the chemical reaction $C_3H_6O + C_2H_6O_2 \longrightarrow C_5H_{10}O_2$ (+ H_2O), the first level of description contains the atoms $(C, C, O, ...)$ and bonds (C=O,C-C, ...). The second level of description contains the molecules $(C_3H_6O, C_5H_{10}O_2, ...)$ and the last level gathers the information about the reaction, *i.e.*, $C_3H_6O + C_2H_6O_2 \longrightarrow C_5H_{10}O_2$ $(+\ H_2O)$. The information contained in the different objects is either predetermined and stored in the organic reaction database (we will come back to database preprocessing later) or derived from an elementary chemical analysis algorithm which allows to extract the different objects associated with a molecular (sub)structure drawn on the computer screen.

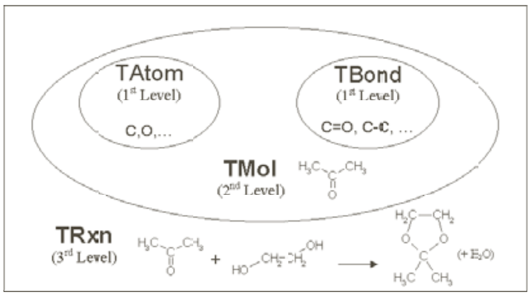


Fig. 1. Presentation of the three level of description for the chemical reaction C_3H_6O $+\ C_2H_6O_2 \longrightarrow C_5H_{10}O_2$ $(+H_2O)$.

3 Screening Descriptors

As mentioned before, the selection of the screening descriptors is of prime importance. Too strong a discrimination between the individual components of a database need to be avoided, otherwise the number of components in each subset is too small, and the number of descriptors that are needed to partition the database is close to the number of individual components in this database. The computer management of all these descriptors thus becomes heavy and the efficiency of the search algorithm strongly decreases. In conclusion, a good screening descriptor must have a discrimination power neither too small nor too large for the database in use.

In order to optimize the screening step, we propose the application of a succession of different searching criteria. The subset of potential candidates is then determined as the intersection depending on all the previously obtained sets. This approach requires that the selected descriptors are lineary independent, in other words, the subsets of potential candidate generated with these descriptors are as distinct as possible.

The discrimination power of a screening descriptor is dependent on the database in use. For example, a descriptor based on the chemical bond type only will not be interesting when working with a database mainly composed of alkane type molecules, *i.e.*, molecules whose atoms are connected through a unique kind of chemical bonds, single bonds (*i.e.*, $C-O$, $C-C$, $C-H$, ...). The screening descriptors must thus ideally be based on the principal characteristics present among the chemical structures of the database.

3.1 Molecular Descriptors

For many computer chemical applications, and particularly for the storage and retrieval in large structural databases, the structure of a molecule can be conveniently represented in terms of a graph. A graph, denoted by G, is a dimensionless mathematical object representing a set of points, called *vertices*, and the way they are connected. The connections are called *edges*. Formally, neither vertices nor edges have a physical meaning. The physical (or chemical) meaning depends only on the problem the graph is representing. In chemistry applications, when a graph is used to depict a molecular structure, *e.g.*, a planar developed formula, it is called a *molecular graph*. A comparaison between a planar developed formula and a molecular graph, applied to the reserpine structure, is presented in Figure 2. The most basic descriptors of molecular graphs are based on the features of the vertices and the edges (number of vertices, egde multiplicity, ...).

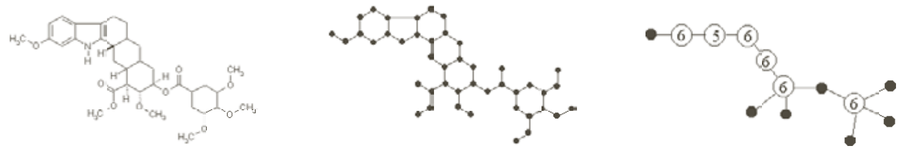
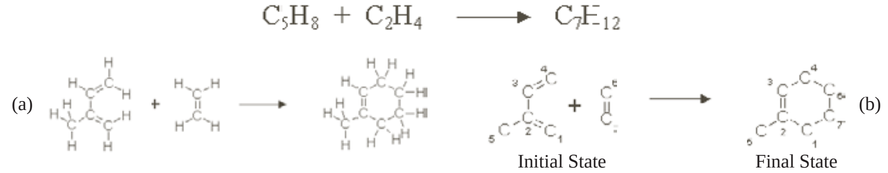


Fig. 2. Planar molecular structure, G, and $RG^{(cycle)}$ of the reserpine molecule. The open circles of $RG^{(cycle)}$ symbolize the vertices containing cyclical information, and the black circles, the vertices containing the non-cyclical information. Labels in the open circles correspond to the size of the reduced rings.

Several different reduced representations of G can be generated. Reduction involves the merging of certain features of a chemical structure into the nodes of a reduced graph. Its purpose is to bring a homeomorphic mapping from the structure onto a simpler graph, resulting, in general, in a smaller number of nodes than in the original graph [3]. As our aim is to obtain a good molecular structure descriptor for the efficient and reliable screening step of structures and substructures present in standard organic reaction databases, and as cyclic structures are widely present in organic chemistry, an obvious first idea is to work with reduced graphs (RG) based on the *cyclic* information, $RG^{(cycle)}$, contained in the molecular structures. We did not find any completely satisfying ring set definition corresponding to our needs in the literature; we have thus chosen to develop a new set of rings as well as the corresponding reduced graph [4]. This new set does not intend to compete with others in terms of strict graph theory considerations or implementation efficiency. We tried to reach a balance between the loss of information (by keeping only certain cycles), and the ease of substructure retrieval. In order to construct this set, we start with the computation of the Smallest Cycles at Edges, by searching the smallest cycles for each edge. In a second stage, the internal rings are included to the set of rings. The obtained ring set is then called the Set of Smallest Cycles at Edges. The reduced graph based on this last set, the $RG^{(cycle)}$, is the junction of two subgraphs, the Graph of Smallest Cycles at Edges, $GSCE$, and the Graph of Acyclic Subtree, GAS. As an example, the $RG^{(cycle)}$ of the molecular structure of reserpine is also depicted in Figure 2.

3.2 Reaction Descriptors

To describe an organic reaction, it is important to find the reaction pattern. For this operation, the elementary principles of connectivity matrix analysis is used. In our case, this matrix contains in its $[i, j]$ cells the multiplicity of the bond joining the i^{th} and j^{th} atoms of the studied molecular structure, *i.e.*, 1 for a single bond, 2 for a double bond, ... As shown in Figure 3, the difference between the final state of the reaction, described in our algorithm by a simple connectivity matrix of the reaction product(s) (E), and the initial reactant, described with the same mathematical tool (B), gives us a new matrix called the reaction matrix (R), where $R = E - B$ [1]. This method is similar to the Dugundji-Ugi approach [5]. The final reaction matrix ($R^{'}$) is then obtained by the suppression of the rows and columns whose all elements equal zero. The resulting $R^{'}$ matrix contains all the transformations occurring during the reaction.



	1	2	3	4	5	6	7
1	0	2	0	0	0	0	0
2	2	0	1	0	1	0	0
3	0	1	0	2	0	0	0
4	0	0	2	0	0	0	0
5	0	1	0	0	0	0	0
6	0	0	0	0	0	0	2
7	0	0	0	0	0	2	0

B

	1	2	3	4	5	6	7
1	0	1	0	0	0	0	1
2	1	0	2	0	1	0	0
3	0	2	0	1	0	0	0
4	0	0	1	0	0	1	0
5	0	1	0	0	0	0	0
6	0	0	0	1	0	0	1
7	1	0	0	0	0	1	0

E

	1	2	3	4	6	7
1	0	-1	0	0	0	1
2	-1	0	1	0	0	0
3	0	1	0	-1	0	0
4	0	0	-1	0	1	0
6	0	0	0	1	0	-1
7	1	0	0	0	-1	0

$R^{'}$

Fig. 3. Reactant connectivity matrix (B), product connectivity matrix (E), and final reaction matrix ($R^{'}$) of a Diels-Alder reaction. (a) Full planar representation of the molecular structure, (b) same as above without the hydrogen atoms.

With this $R^{'}$ matrix, we can determine many features of the reaction, as the number and the nature of the broken and formed bonds, the nature of the chemical functions implied in the transformation, and many other informations. All these informations can be used as reaction descriptors. We originally developed new descriptors based on the eigenvalues of the $R^{'}$ matrix in order to provide a more mathematical representation of a chemical reaction.

4 Discussion and Conclusions

One of our research aims in Computer Aided Organic Synthesis are to built efficient searching algorithms for organic reaction databases. So far, we have

[1] Hydrogen atoms are often omitted in description of organic reactions. They can be easily recovered from the hybridation state of the connected carbon atoms.

adopted a two-step process containing a screening step which involves the computation of a subset of potential candidates, and an ABAS step, used to find the queried reaction. In order to optimize our algorithm, new screening descriptors have been developed in function of the user′s question. We have thus computed several molecular descriptors, such as reduced graphs based on cyclical information, and on the carbon/heteroatom distinction, and several reaction descriptors, based on the information contained in the reaction pattern, such as eigenvalues of the $R^{'}$ matrix, the number and nature of the broken and formed bonds, the nature of the chemical functions implied in the transformation. In order to analyse the discrimination power of the different screening criteria, our strategy was applied to a test database of 540 reactions. According to the results of an histogram analysis, which reports the discrimination power in function of the number of vertices in the reduced graph ($\sharp\{V^{(RG)}\}$) (Fig. 4), three different kinds of behavior can be observed. The discrimination power can be large, sufficient, or too soft. For the last kind (zone C in Fig. 4 left), the reduced graphs under study are clearly not efficient for the structures they represent. Another reducted criteria is thus necessary. The most important interest of the use of a second (or several) type of RG is that the structures that were appearing in the A, B, or C regions with one kind of reduction may appear in other regions with another kind of reduction. Indeed, as shown for example in Figure 5, a structure that appeared with the RG based on cyclical information in area C, *i.e.*, the least discriminating one, may appear in area B, with a higher discrimination power, when another criterion of reduction, the carbon/heteroatom criterion, is used. In our strategy, all the structural and reaction descriptors of the database are computed during a preprocessing step and each histogram is calculated. At each introduction or

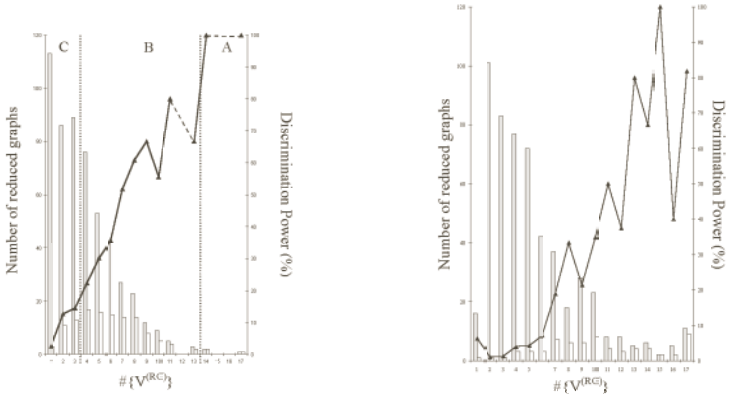


Fig. 4. Histogram showing the action of the reduced graphs (RG) based on the carbon-/heteroatom distinction (left), and based on cyclical information (right), over a test population of 540 molecular structures. Abscissa values are the number of vertices per RG, ordinate values, the number of RG. Gray columns represent the total number of graphs, white columns, only the different graphs. Black triangles symbolize the value of the discrimination power.

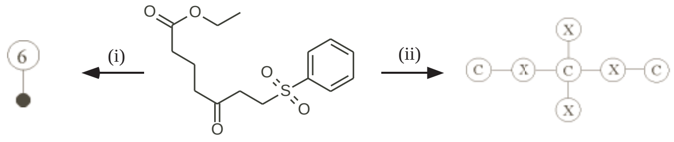


Fig. 5. Comparaison between the RG based on cyclical information (i) and the RG based on the carbon/heteroatom distinction (ii) for the structure of ethyl 4-acetyl-6-phenyl-sulfonyl-hexylate. Labels of the open circles of the $RG^{(cycle)}$ correspond to the size of the reduced rings. Black circles of the $RG^{(cycle)}$ represent non-cyclical information. Letters in open circles of the RG based on carbon/heteroatom reduction symbolize the vertex containing the carbon atom (C label), and the vertex containing heteroatoms (X label).

deletion of an additional chemical reaction, the database changes and thus the histograms may change too. Therefore, the histograms must be updated every time the database changes, but not, of course, every time the user is querying the database.

In conclusion, the building of the histograms associated with each screening descriptor allows to determine the best set of screening criteria, for a given user question and for a given database, and then an optimum screening stage may be established. Indeed, the discrimination power of a criteria is dependent on the database that is used. The use of molecular and reaction screening descriptors is thus a necessary step to allow a fast and efficient search for chemical reaction.

5 Acknowledgments

L.D. thanks Prof. A. Krief, Dr G.M. Downs, and T. Latour for useful discussions, and the “Fonds pour la formation à la Recherche dans l’Industrie et dans l’Agriculture” (FRIA) for his PhD fellowship.

References

1. Corey, E.J., Wipke W.T.: Computer-Assisted Design of Complex Organic Syntheses. Science **166** (1969) 179–192.
2. Bartmann, A., Maier, H., Walkowiak, D., Roth, B., Hicks M.G.: Substructure Searching on Very Large Files by Using Multiple Storage Techniques. J. Chem. Inf. Comput. Sci. **33** (1993) 539–541.
3. Gillet, V.J., Downs, G.M., Ling A., Lynch, M.F., Venkataram, P., Wood, J.V., Dethlefsen W.: Computer Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graphs and Their Applications in Generic Chemical Structure Retrieval. J. Chem. Inf. Comput. Sci. **27** (1987) 126–137.
4. Dury, L., Latour, T., Leherte L., Barberis, F., Vercauteren, D.P.: A New Graph Descriptor for Molecules Containing Cycles. Application as Screening Criterion for Searching Molecular Structures within Large Databases of Organic Compounds. J. Chem. Inf. Comput. Sci., submitted for publication
5. Hippe, Z.: Artificial Intelligence in Chemistry, Structure Elucidation and Simulation of Organic Reactions. *Studies in Physical and Theoretical Chemistry;* Elsevier, PWN Polish Scientific Publishers (1991), vol. 73, 153–183.