A New Graph Descriptor for Molecules Containing Cycles. Application as Screening Criterion for Searching Molecular Structures within Large Databases of Organic Compounds

Laurent Dury,*,† Thibaud Latour,† Laurence Leherte,† Frédéric Barberis,†,‡ and Daniel P. Vercauteren†

Laboratoire de Physico-Chimie Informatique and Laboratoire de Chimie Organique de Synthèse, Facultés Universitaires Notre-Dame de la Paix, Rue de Bruxelles 61, B-5000 Namur, Belgium

Received October 17, 2000

The search of molecular structures inside a large database of chemical compounds is a critical step for many computer programs used in several domains of chemistry. During the last years, the size of many chemical databases has dramatically increased, hence in the meantime, search engines needed to be more and more powerful. The speed and the efficiency of screening processes of the chemical compounds are thus essential. Looking forward for algorithms dedicated to structure and substructure search, we have developed a new graph descriptor for structures containing cycles in order to find efficient indexation and classification criteria of molecular structures. This graph descriptor can be used as a screening criteria for structure and substructure search in large databases of organic compounds.

1. INTRODUCTION

Consulting chemical information stored in various types of databases through computer means is increasingly important. During the last two decades, organic chemists, for example, started to use computer programs in organic synthesis. Computer Aided Organic Synthesis programs were developed to discover strategies for the preparation of target molecules. This method was first demonstrated in 1969 by Corey and Wipke[1] and was based on Corey's original concept of retrosynthetic analysis.

Organic chemists are also more and more involved in consulting reaction databases to find information about reactions and/or about reactants. Other scientists working for example in the biopharmacological or fine chemistry domains often need to compare the physical chemistry properties of new molecular structures with data already present in the literature. Therefore, they consult crystallographic structural databases or spectral databases.

About 20 years ago, new types of reaction databases were developed on computer platforms. These were structure-based reaction indexing systems which focus on reaction centers. In addition to molecular structures and literature references, these databases usually contain reaction conditions, reaction efficiency, catalyst agents, and many other information. Commercial versions of these types of systems are for example ORAC (Organic Reactions Accessed by Computer), REACCS (REaction ACCess System), or SYNLIB (SYNthesis LIBrary).

Since 1990, a number of new reaction databases have appeared[2] such as CASREACT,[3] ChemInform,[4] ChemReact, and Beilstein CrossFire,[5,6] and, more importantly, accession to the various reaction database servers was significantly improved. The size of all databases, in terms of molecular structures as well as of reactions, is increasing dramatically every year. The efficiency of search algorithms implemented within these databases is thus crucial. Hence the computer-aided database managers have become useful tools for many research laboratories in industry and academia.

Optimization of search algorithms generally consists of two steps. First the *screening* determines, with the help of a set of descriptors based on the user's question, a subset of potential candidates. In a second time, all components of the subset are compared with the user question to find those that are searched for. This second step is usually called ABAS (Atom by Atom Search).[7] The ABAS 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 accuracy of the choice determines the performance of the database search.

Nevertheless, too strong a discrimination between the individual components of the database need to be avoided, otherwise the number of components in each subset would be too small, and the number of descriptors necessary to decompose the database would tend to approach the number of components in this database. The computer management of all these descriptors would then become too heavy, and the efficiency of the search algorithm would decrease dramatically. To avoid this problem, one should represent the target with an appropriate level of information generalization. The result is a searching criteria which will depend on the paradigm of the reduction (e.g., cycle, functional groups, hydrocarbon groups, ...). In addition, a good com-

* Corresponding author phone: +32-81-734534; fax: +32-81-724530; e-mail: laurent.dury@fundp.ac.be. FRIA Fellow.
† Laboratoire de Physico-Chimie Informatique.
‡ Laboratoire de Chimie Organique de Synthèse.

10.1021/ci000401y CCC: $20.00
Published on Web 09/29/2001

promise is to perform the screening step using only these criteria for which a good discrimination power has been determined during a preprocessing phase of the database (DB). In other words, the searching criteria must fit to the DB diversity in terms of screening efficiency. This topic will be discussed at the end of this paper. For example, a descriptor based on the sole bond type paradigm will not be efficient when working with a DB mainly composed of alkane type molecules.

In conclusion, a good screening descriptor for a given quest entry must have a discrimination power neither too small nor too large relative to the DB in use. The screening descriptors must ideally be based on the principal characteristics that are present among the chemical structures as well as on their distributions. One important kind of descriptor is the information concerning cycles. Indeed, rings have a crucial importance in chemistry and have been reported in *Chemical Abstract* since 1907.[8] Information contained in rings is a large part of the structural topology used to identify and characterize molecular structures.

In this paper, we will first recall some basics about molecular graph representation and ring information. After a quick description of several existing sets of rings, a new ring set is proposed, and with this set, a new reduced graph based on the ring information paradigm (the $RG^{(\mathrm{cycle})}$) is defined. And finally, in part 5, as our goal is to develop a powerful structure and substructure searching algorithm, we detail the study of the discrimination power of the set of searching criteria (e.g., number and type of super-atoms, ...) contained in our new reduced graph over a molecular structure test population.

2. MOLECULAR GRAPH REPRESENTATION

For many chemical applications, and particularly for storage and retrieval in large databases, the structure of a molecule can be conveniently represented in the form 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 (or are not) 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, we call it a *molecular graph*.

A molecular graph G is constituted by two distinct sets: E, the set of N_e edges e_i representing the bonds; and V, the set of N_v vertices v_i representing the heavy atoms (hydrogens are usually not taken into account). Hence, $G = \{E, V\}$ with $E = \{e_1 \cdots e_{N_e}\}$ and $V = \{v_1 \cdots v_{N_v}\}$. An edge $e = \{v_i, v_j\}$ is formed by a connected pair of vertices. Two vertices connected by an edge are *adjacent*. In the present context, the molecular graphs are *nonoriented* and *nonweighted*, that is, $e = \{v_i, v_j\} = \{v_j, v_i\}$, and all edges and vertices have the same unit weight, respectively. The graph vertices and edges may be *labeled*. However, we shall not use this feature in the present discussion. Rather than being connected by several edges, a multiple bond like a π-bond will be represented by a single edge with adequate multiplicity (2, for instance). This is done to avoid meaningless two-membered cyclic patterns in the representation.

To keep the following discussion clear and concise, we now recall some basic definitions. The *degree* of a vertex v_i, denoted $d(v_i)$, is the number of edges involving v_i. Thus, $d(v_i) = \#\{e: v_i \in e\}$. A *node* consists of the union of a vertex and all edges connecting its adjacent neighbors: $n_i = v_i \cup \{e: v_i \in e\}$. A *subgraph* G' is obtained by removing one or several nodes from G. A *walk* is formed by traversing the graph through a succession of *adjacent* vertices between two *endpoints*. A *cycle* is a closed walk with no repeated edges and in which one and only one vertex is repeated; the endpoint. The size of a cycle is denoted by σ. In the discussion, we shall not make any distinction between rings and cycles, these two words being considered as synonymous.

Usually, a molecular graph is a *connected* graph. In such a graph, every two distinct vertices are joined by a path, otherwise it is *disconnected*. The *component* of G is the maximal connected subgraph of G. If the graph is disconnected, there is more than one component. Hence, a water dimer would be represented by a disconnected graph with two components, provided that hydrogen bonds are not represented by edges. Similarly, catenanes would also be represented by a disconnected graph with n components.

Most chemical graphs can be drawn in two dimensions such that no edges cross (catenanes are exceptions to this general property), in which case the graph is said to be embeddable on a plane. A planar embedding defines regions (which have no edges or vertices crossing them) and nonregions (in other cases). The region defined by the outer boundary of a graph is the infinite region; all other regions are finite regions. Finite and infinite regions are always interchangeable by redrawing the graph. Following the terminology introduced by Downs,[9] rings can be classified in terms of faces of a planar embedding. A simple-cycle is a cycle in which no pair of vertices is joined by an edge not in the cycle. A region that is a simple-cycle is a simple-face. A nonregion that is a simple-cycle is a cut-face. If a cut-face is the smallest simple-cycle associated with at least one of its edges, then it is a primary cut-face; otherwise, if at least one of its edges is associated with a simple-cycle of the same size, then it is a secondary cut-face; otherwise it is a tertiary cut-face. A planar simple-cycle is the two dimensional region, delimited by the edges, which includes no vertex.

3. RING INFORMATION

When looking at molecular graphs, one can distinguish two parts in G regarding cycle information. We denote $\mathcal{G}$, the "cyclic" subgraph of G. $\mathcal{G}$ is obtained by recurrently removing all nodes for which the vertices have $d(v_i) \leq 1$ until no such node subsists in $\mathcal{G}$. The subgraph $\mathcal{G}$ is formed by the set $\mathcal{E}$ of the $\mathcal{N}_e$ edges involved in cycles, and the set $\mathcal{V}$ of the $\mathcal{N}_v$ vertices involved in cycles. In addition, all nodes which cannot be endpoint of any cycle are also removed. The removed nodes forms the "acyclic' subgraph g of G. Hence, with $\mathcal{G} = \{\mathcal{E}, \mathcal{V}\}$:

$$g = G - \mathcal{G} = \{E - \mathcal{E}, V - \mathcal{V}\} \tag{1}$$

Figure 1 depicts the G, $\mathcal{G}$, and g graphs corresponding to the molecular structure of reserpine.

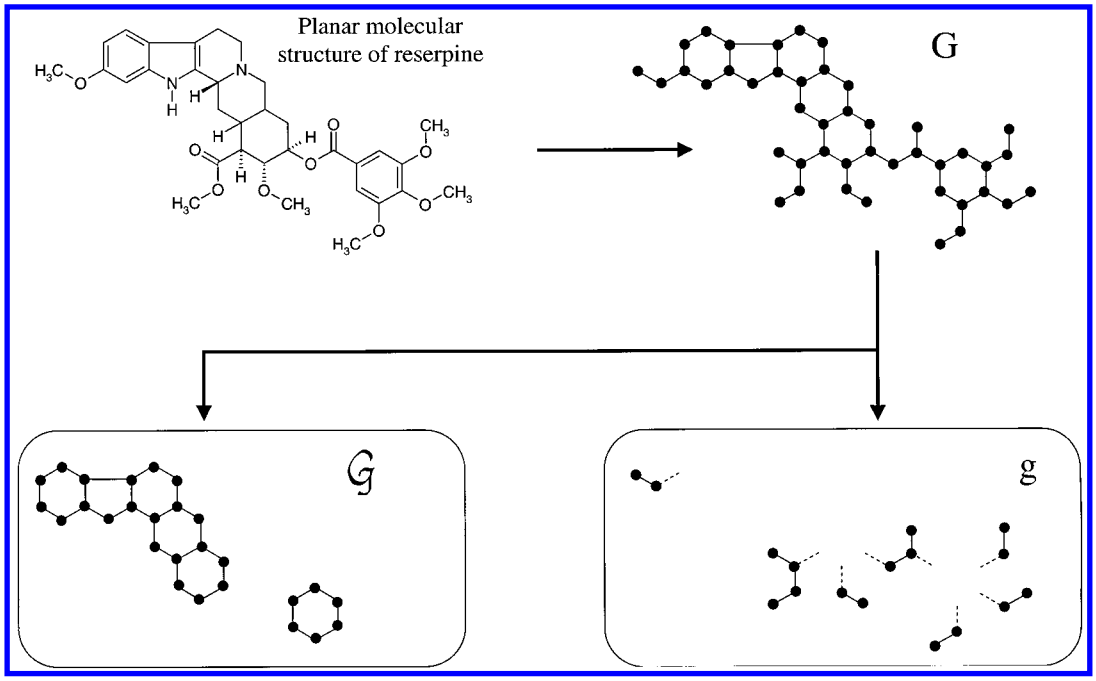

Figure 1. The G, $\mathcal{G}$, and g graphs corresponding to the planar molecular structure of reserpine.

Any set of rings can be characterized by its σ-sequence, i.e., the sequence of sizes of all rings of the set in an ascending order. A σ-sequence A is smaller than a σ-sequence B if

$$(i)\ \#\{A\} = \#\{B\}$$

$$(ii)\ \textit{the first } k-1 \textit{ elements of } A \textit{ and } B \textit{ are equal}$$

$$(iii)\ \textit{the } k_{\text{th}} \textit{ element of } A \textit{ is smaller than the } k_{\text{th}} \textit{ element of } B \quad (2)$$

A cycle $\rho(\sigma)$ is the *smallest cycle at edge* e_i if its size is the smallest of all cycles containing e_i, we note it $\rho^{\dagger}(e_i,\sigma)$.

The ring information within a molecular structure consists of the sets which describe all or some of the molecular rings as well as the relations between them. Many kinds of ring sets have already been described in the literature. The most basic of them is the set of all possible rings, Ω. However, the description of a molecule in terms of all its constituting rings may be difficult to handle, especially because of the combinatorial explosion of the size of this type of set. Therefore other sets are usually used, e.g., the Smallest Set of Smallest Rings (SSSR), the Essential Set of Essential Rings (ESER), the Extended Set of Smallest Rings (ESSR), and many others. These new sets are all subsets of Ω. A useful compilation discussing all subsets listed above as well as many others (β-rings, K-rings, ...) can be found in a paper by Downs et al.[10]

Among the above cited sets, the SSSR is still the most important and widely used. It contains the μ smallest simple-cycles in a molecular structure, expressed as

$$\mu = \mathcal{N}_e - \mathcal{N}_v + C \quad (3)$$

where C is the number of components in $\mathcal{G}$. None of the SSSR rings can be represented as a linear combination of the others, hence, SSSR rings form a basis of the ring space.

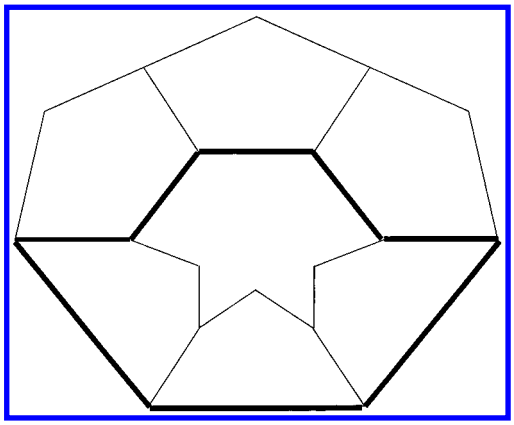

Figure 2. Graph schematizing an hypothetical complex molecular structure with a missing planar simple-cycle (the internal nine-membered ring) in the SSSR.

In addition, the SSSR has the smallest σ-sequence among all possible bases in the ring space.

As our goal is to determine the most efficient descriptors for a rapid structure or substructure search in a two-dimensional molecular structure database, we chose to work with a set of rings which describes all the simple rings of the planar projection of each molecular structure. The SSSR cannot be efficiently used for that purpose because in some complex molecular structures, this set does not allow the description of all chemically meaningful rings. Indeed, a "planar simple-cycle" of a cyclic substructure may be missing in the SSSR. The structure presented in Figure 2 (from ref 10) illustrates one example of such a situation. Its SSSR contains a eight-membered ring (tertiary cut-face), in bold on the figure, instead of the expected nine-membered central ring (planar simple-cycle).

Moreover, if one wishes to use the ring information as a screening criterion for substructure search, the ideal (and strict guarantee for exactness) solution would be to store the Ω set of each structure in the database, and check if the Ω set of the searched substructure is a subset of the Ω set of

each database entry. This trivial solution is of course in practice quite difficult to handle. However, we *need* to know during the screening process, whatever the method, all members of the target structure Ω set (say a database entry), otherwise, the result of the search will miss some correct answers. In theory, the SSSR definition allows us to combine any members of the SSSR to obtain any member of the Ω set. Unfortunately, this operation, regarding our efficiency objective, may rapidly be cumbersome, especially if we keep in mind that it has to be done for each database entry. Therefore, we have to find a good balance between the amount of information stored and the simplicity in the combination of cycles.

To overcome this problem, some modifications to the SSSR were introduced. The simplest modification consists of the addition of some extra cycles belonging to Ω, the SSSR, to form the so-called SSSR$^+$. Also, the Essential Set of Essential Rings (ESER), closely related to the SSSR and originaly defined by Fujita,[11] was designed to describe the imaginary transition structures corresponding to the reaction-site changes during an organic reaction sequence. It consists of the union of a series of simple cycles selected according to a so-called "nondependency" criterion (see Fujita's original paper for full details). Finally, the Extended Set of Smallest Rings (ESSR), defined by Downs et al.,[10] is the unique set of rings which contains all simple-faces, all primary cut-faces, and all secondary cut-faces. This set contains all the rings necessary to our work, i.e., all the "chemically meaningful cycles", but also unfortunately some other less relevant ones for our rapid database screening purpose.

Hence, as we did not find any completely satisfying ring set definition corresponding to our needs, we have chosen to develop a new set of rings as well as the corresponding reduced graph. This new set does not intend to compete with the others in terms of strict graph theory considerations or implementation efficiency. However, since our aim is to obtain a good molecular structure descriptor for efficient and reliable database screening, we tried to reach a balance between the loss of information (by keeping only certain cycles) and the ease of substructure retrieval.

4. REDUCED GRAPHS BASED ON RING INFORMATION

Starting from G, the molecular graph, we can build some reduced graphs (RG). This reduction involves the merging of certain features of the chemical structures in the nodes of the reduced graphs according to a paradigm (say, a property, a characteristic). Its purpose is to bring about a homeomorphic mapping from the structure onto some simpler graph, resulting, in general, in a smaller number of nodes than in the original graph.[12] A paradigm used to build these reduced graphs is the grouping together in terms of ring and nonring components. An RG will thus be obtained by the connection of two subgraphs: one containing the cyclic information (the Graph of Smallest Cycles at Edges) and another containing the noncyclic information (the Graph of Acyclic Subtrees). More generally the two components will be, on one hand, the subgraphs for which the paradigm requirement is verified, and, on the other hand, the remaining subgraphs for which the requirements are not met.

4.1. The Graph of Smallest Cycles at Edges. Let $G = \{E, V\}$ be a molecular graph and $\mathcal{G} = \{\mathcal{C}, \mathcal{V}\}$, its "cyclic" subgraph as defined above. We recurrently construct the *SSCE* (Set of Smallest Cycles at Edges). To do so, we compute the *k*th intermediate set: $SCE^{(k)}$. The $SCE^{(k)}$ is the set of the $\rho^{\dagger}(e_i^{(k)}, \sigma)$, for all edges $e_i^{(k)}$. Hence

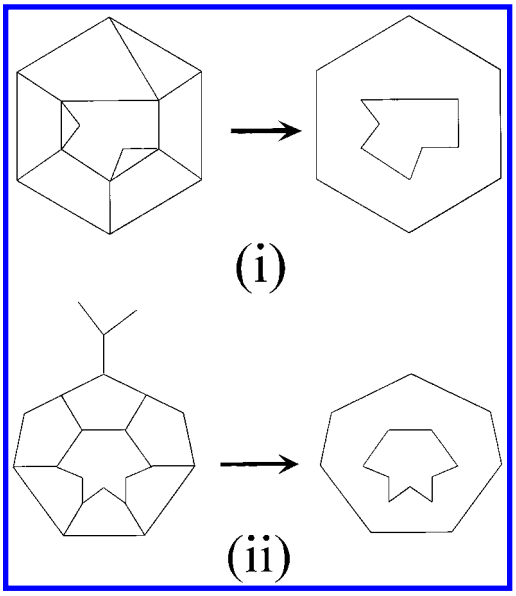

Figure 3. G (right) and $\mathcal{G}^{(2)}$ (left) graphs of two hypothetical complex molecular structures. Structure (i) illustrates human optical misunderstanding. Indeed the enveloppe ring (i.e., the largest cycle) is drawn inside the graph picture. Molecular structure (ii) corresponds to the particular case where the sizes of the internal and envelope rings are of equal value.

$$SCE^{(k)} = \bigcup_{i=1}^{\mathcal{N}_e^{(k)}} \rho^{\dagger}(e_i^{(k)}, \sigma) \tag{4}$$

Its cardinality $\#\{SCE^{(k)}\}$ is $\mathcal{N}_c^{(k)}$, and sce_i is an element of the $SCE^{(k)}$. After the first run ($k = 1$), there might be some important cycles missing in the $SCE^{(1)}$, such as some faces in convex polyhedrons or simple elementary cycles which are not the smallest cycles of any edge. Therefore, to find those remaining cycles, i.e., the next *k*th iteration, we clean from $\mathcal{C}$ all the edges shared by at least two sce_i cycles. Thus

$$\mathcal{C}^{(k+1)} = \{e | (e \in sce_i) \wedge (e \notin \bigcup_{j=1, j\neq 1}^{\mathcal{N}_c^{(k)}-1} sce_j, \forall i, j)\} \tag{5}$$

We also have that $\mathcal{N}_e^{(k+1)} = \#\{\mathcal{C}^{(k+1)}\}$ and $e_i^{(k+1)} \in \mathcal{C}^{(k+1)}$. Hence, the remaining graph is $\mathcal{G}^{(k+1)} = \{\mathcal{C}^{(k+1)}, \mathcal{V}^{(k+1)}\}$ with $\mathcal{V}^{(k+1)} = \mathcal{V}^1$. Indeed, during this iteration stage we only work with edges. Therefore, it would be computationally useless to clean the $\mathcal{V}^{(k+1)}$ set by applying the recursive Cyclic Subgraph Algorithm we presented earlier at the beginning of Section 3. Of course, it would have been more elegant to process the $(k + 1)$th iteration with the cyclic subgraph of the cyclic subgraph, at the *k*th level. But this is only a matter of formalism detail we shall not consider here. Figure 3 compares the G and $\mathcal{G}^{(k+1)}$ graphs for two examples. In the first structure (i), $\#\{SCE^{(1)}\}$ is 8. The sce_i are five four-membered rings and three three-membered rings. The $\#\{SCE^{(2)}\}$ is 2, the internal ring (region delimited by the rings of the $SCE^{(1)}$) and the envelope ring, the infinite region. By definition, the envelope ring is the largest ring (the seven-membered ring in our case). Contrary to the first sight impression the envelope ring is not the external six-

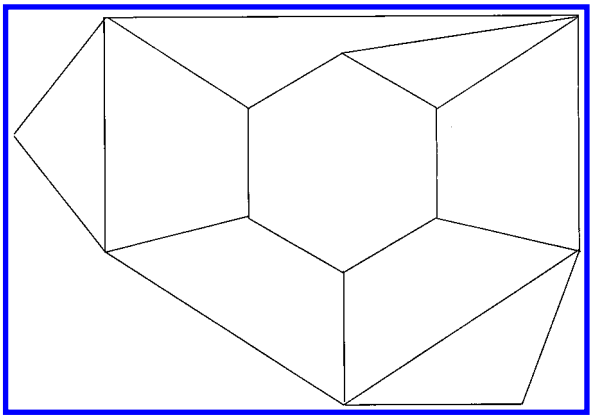

Figure 4. Planar molecular structure (i) of Figure 3 represented after a change of embedding.

membered ring in the drawing but the inner seven-membered ring. Since this ring is not a planar simple-cycle, we chose to remove it from $\mathcal{G}^{(2)}$ if its size differs from one of the internal rings. Indeed, in this particular case, it is not possible to differentiate between the two rings, the suppression of one of them leads to a nonunique set of rings and next to a nonunique reduced graph. The second structure (ii) shown in Figure 3 depicts such a case; the suppression of one of the rings contained in $\mathcal{G}^{(2)}$ leads to two different reduced graphs (we will come back to this point later).

To summarize, if more than one planar embedding is possible and if the size of the envelope ring differs from the size of the internal ring, then $\mathcal{G}^{(2)}$ is the set of regions from the embedding with the largest infinite region.

The complementary $SCE^{(k+1)}$ is obtained by the same rule as in relation (4), that is

$$SCE^{(k+1)} = \bigcup_{i=1}^{\mathcal{N}_e^{(k+1)}} \rho^{\dagger}(e_i^{(k+1)},\sigma) \qquad (6)$$

We then repeat the operations described in relations (5) and (6) q times until $SCE^{(q)} = \emptyset$. Finally, we group all $SCE^{(k)}$'s ($k = 1...q$) to get the final set of cycles (7). To facilitate the understanding of this paper, we will give to our final set of cycles the name of Set of Smallest Cycles at Edges, SSCE. Of course, this only means that the elements of the set are the smallest cycles at the edges *within one of each k construction*.

$$\mathrm{SSCE} = \bigcup_{k=1}^{q} SCE^{\mathrm{k}} \qquad (7)$$

In the second structure of Figure 3, the SSCE contains the six five-membered rings and the two nine-membered rings; it is then similar to the ESSR. On the other hand, in the first structure, the ESSR is {3, 3, 3, 4, 4, 4, 4, 4, 5, 6}, while the SSCE is {3, 3, 3, 4, 4, 4, 4, 4, 6}. The five-membered ring is not a smallest cycles at edges. Note also that the seven-membered ring found in the $SCE^{(2)}$ is not kept in the *SSCE*. It is the envelope ring of the structure (Figure 4). To summarize, as show in Figure 5 for four different cyclic structures, the SSCE may be either similar to the SSSR set or to the ESSR set or may be different. The *SSCE* contains the planar simple-cycle and the primary and secondary cut-face. Indeed, as shown in Figure 6,[9] the three-membered primary cut-face (left) and the four-membered

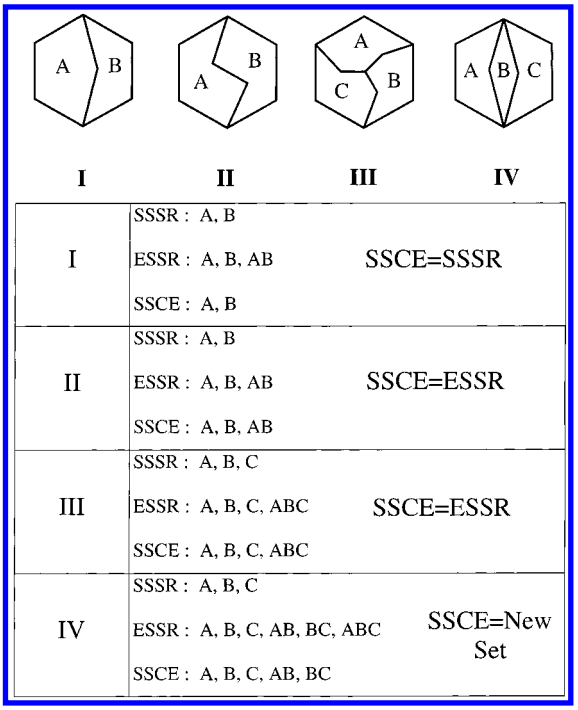

Figure 5. Comparison of the *SSSR*, *ESSR*, and *SSCE* for some molecular structures.

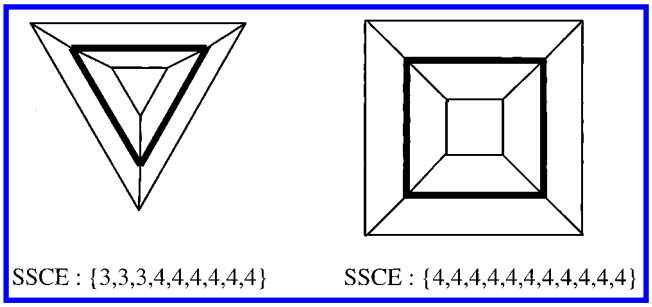

Figure 6. Graphs schematizing hypothetical complex molecular structures containing a primary cut-face (left) or a secondary cut-face (right).

secondary cut-face (right), both in bold in the figure, are parts of the *SSCE*. Unlike the ESSR, the envelope ring is kept in the *SSCE* only if the graph has an internal ring whose size is the same as the envelope ring size (we will come back on this point later).

Once this operation is done, we are able to create the *GSCE* (Graph of Smallest Cycles at Edges), a new graph in which the vertices are the elements of the ring set and the edges represent either the shared $\mathcal{V}$ vertices (spiro cycle connectivity) or the shared $\mathcal{E}$ edge(s) (edge cycle connectivity) with a label depending on the number of shared edges. The *GSCE* of the reserpine molecule presented in Figure 1 is displayed in Figure 7. The label of each vertex of the *GSCE* is function of the size of the ring describing the vertex. To each edge, we associate a multiplicity value symbolizing the number of bonds shared between two reduced rings (Figure 8). A single line corresponds to the sharing of one edge, a double line, two edges, and a triple line, three edges or more. The dash edge symbolizes a simple junction between two rings, in chemical words, a bond joining two cyclic groups.

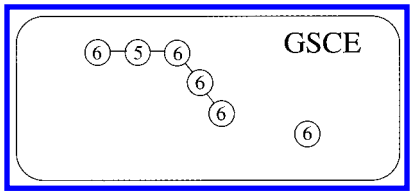

Figure 7. *GSCE* of the reserpine molecule. The open circles symbolize the vertices. The label in the open circles correspond to the size of the rings reduced in each of the *GSCE* vertex.

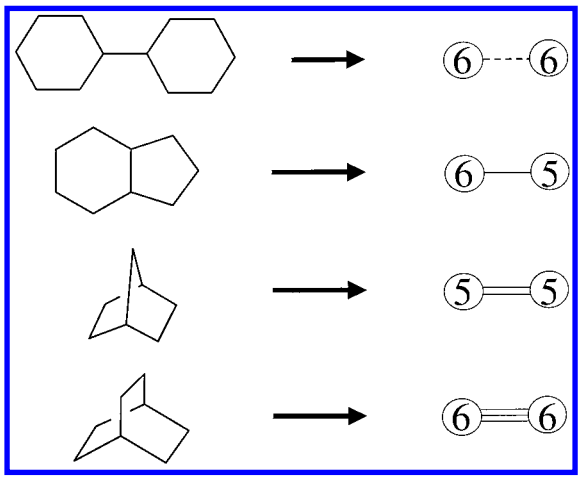

Figure 8. Presentation of the different kinds of multiplicity of the edges of the *GSCE*.

4.2. The Graph of Acyclic Subtrees. Having reduced the graph information in terms of cycles in the *GSCE*, we now focus on the "acyclic" information in *G*, i.e, the information which does not fulfill the cyclic paradigm requirement. The reduction is here very simple and uses the *g* subgraph defined in section 3. There are in *g* as many component as acyclic subtrees contained in *G* and each of these components is reduced to a single super-vertex. Doing so, we obtain the *GAS* (Graph of Acyclic Subtrees). The *GAS* is therefore by definition a totally disconnected graph with each vertex having a degree of zero. Figure 9 compares the *g* and *GAS* of the reserpine molecule presented before. We can observe that each component of the *g* graph is reduced to one super-vertex (black circles in the figure). When the *GAS* is constructed, our algorithm keeps all the information (nature, charge, ...) about each atom contained in the *g* component. We will give some insight in this hierarchical ordering in an other forthcoming paper.

4.3. The Cyclic Information Reduced Graph: $RG^{(cycle)}$. Finally, to express the molecular structure descriptor in terms of a reduced graph containing the cyclic information, we merged both *GSCE* and *GAS* graphs. The connections between the cycle, like super-vertices of *GSCE* and the noncycle, like super-vertices of *GAS*, are done using a very simple rule:

If there exists one edge in G connecting a vertex belonging to a GSCE element and a vertex belonging to a GAS element, these two elements are adjacent in $RG^{(cycle)}$.

Because of the definition of $\mathcal{G}$, and also the one of *g*, there is no more than one edge connecting a *GSCE* element and a *GAS* element. In addition, if *G* is a connected graph, $RG^{(cycle)}$ is also a connected graph. Coincidentally, the number of components in *G* is equal to the number of components in $RG^{(cycle)}$. Figure 10 presents the $RG^{(cycle)}$ of the reserpine molecule.

We can now understand the problem arising when the envelope ring and the internal ring have the same size. Indeed, if one of the two similar rings is chosen randomly, it will not change the *GSCE*, but it will certainly modify the connections between the vertices of the GSCE and the GAS. It would then lead to two possible $RG^{(cycle)}$'s and an ambiguous working definition of $RG^{(cycle)}$. The two reduced graphs obtained with the second structure (ii) in Figure 3 are represented in Figure 11. The open circle symbolizes the vertices of the *GSCE* graph, and the black circles represent the vertex of the *GAS* graph. The label in the open circles corresponds to the size of the rings reduced in each *GSCE* vertex. The multiplicity of the edges symbolizes the

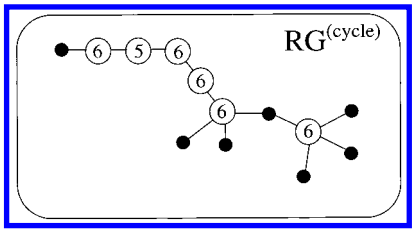

Figure 10. $RG^{(cycle)}$ of the reserpine molecule. The open circles symbolize the vertex of the *GSCE* graph, and the black circles, the vertex of the *GAS* graph. The label in the open circles correspond to the size of the rings reduced in each of the *GSCE* vertex.

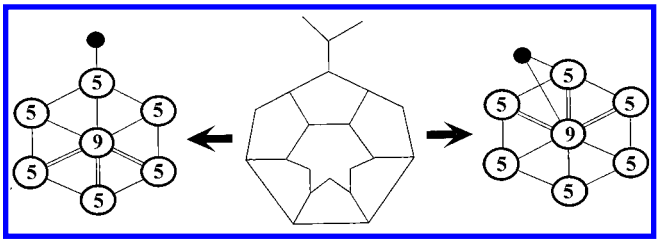

Figure 11. Comparison of the two different $RG^{(cycle)}$'s built from the hypothetical molecular structure (ii) of Figure 3 in which the size of internal and envelope rings are the same. The open circles symbolize the vertex of the *GSCE* graph, and the black circles the vertex of the *GAS* graph. The label in the open circles corresponds to the size of the rings reduced in each of the *GSCE* vertex. The multiplicity of the edges symbolizes the number of bonds shared between two reduced rings.

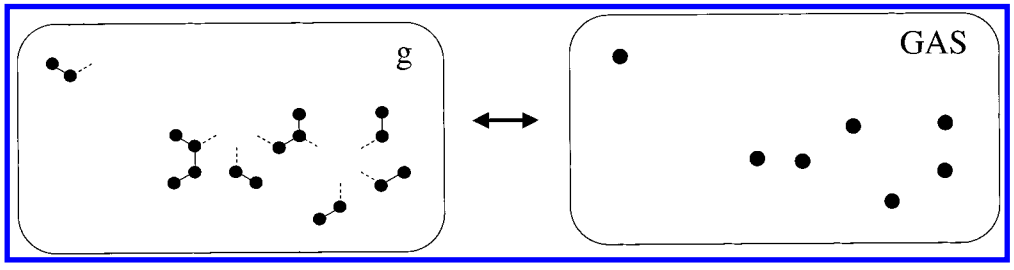

Figure 9. Comparaison of the *g* (left) and *GAS* (right) graphs applied to the reserpine molecule. The black circles represent the vertices.

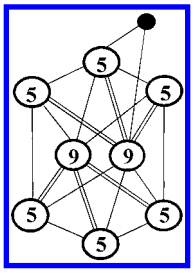

Figure 12. $RG^{(cycle)}$ of the hypothetical molecular structure (ii) of Figure 3 obtained by adding the envelope ring. The open circles symbolize the vertices of the *GSCE* graph, and the black circles represent the vertex of the *GAS* graph. The label in the open circles corresponds to the size of the rings reduced in each of the *GSCE* vertex. The multiplicity of the edges symbolizes the number of bonds shared between two reduced rings.

number of bonds shared between two reduced rings. We show then that the envelope ring of this structure must stay in $\mathcal{G}^{(2)}$. The $RG^{(cycle)}$ obtained with this correction is presented in Figure 12. Importantly, this reduced graph is invariant with the change of embedding of the G graph.

5. STRUCTURE SEARCH USING THE $RG^{(CYCLE)}$ CONCEPT IN A MOLECULAR DATABASE.

As the $RG^{(cycle)}$ contains a lot of information about the molecular graph, it can be used as a descriptor for the screening step. The proposed original approach was applied to a test database (DB) of 540 organic molecular structures in order to determine its discrimination power. The $RG^{(cycle)}$'s of all the DB entries have been constructed and then compared to each other to remove redundant identical reduced graphs. This operation corresponds to the selection of archetype entries present in the DB according to the actual paradigm. The results of our analysis are reported in Figure 13. The discrimination power is evaluated by the ratio of the sum of all the graphs having the same number of vertices over the sum of the graphs having this same given number of vertices but presenting a different connectivity between them. The discrimination power is symbolized by the black triangles.

The obtained results allow us to distinguish three different categories of discrimination power. In area A, the power of discrimination is maximum, i.e., 100%. To each structure corresponds one and only one reduced graph. Then, if the $RG^{(cycle)}$ is built, the queried structure is found. Indeed, the screening step will give a subset of potential candidates which have a cardinality equal to one. The only potential candidate is the solution. The searching process is made in one step, the ABAS step is thus useless. In practice, a such discrimination power value does not allow an optimum search because the number of descriptors thus equals the population of the database. But with the population of structures we have analyzed, the structures that have this kind of $RG^{(cycle)}$ are rare, and the problem does not matter. The molecular structures present in this area contain at least 14 cyclic or acyclic regions. An example of this kind of molecular structure is shown in Figure 14.

In area B, the discrimination power is lower, between 80 and 20%; it is still efficient enough for a good screening step. This area is the optimum region if we compare the number of the structures studied with the average value of the discrimination power.

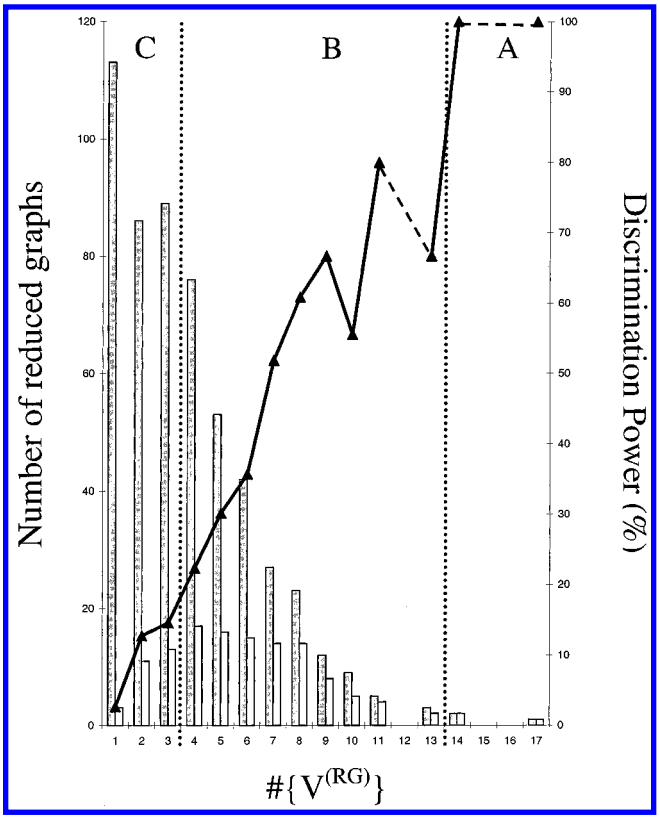

Figure 13. Histogram showing the action of the $RG^{(cycle)}$ concept over a test population of 540 molecular structures. Abscissa values represent the $\#\{V^{(RG)}\}$, ordinate values, the number of RG. Gray columns represent the total number of graphs, white columns, only the number of different graphs paradigm or archetypes. Black triangles symbolize the value of the discrimination power.

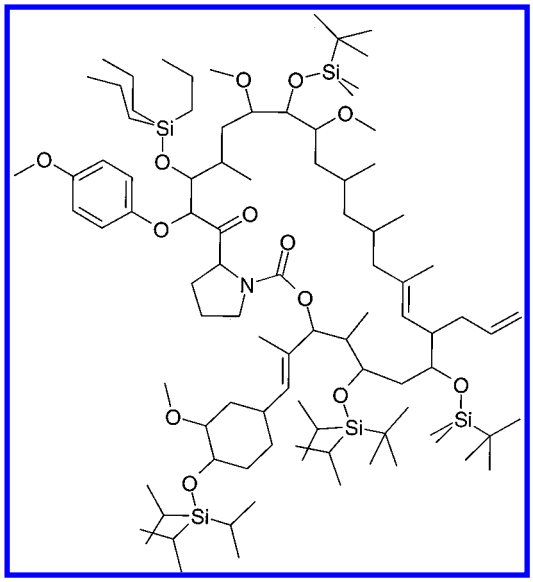

Figure 14. Example of a molecular structure located in the area A of the histogram presented in Figure 13.

In the last area C, the discrimination power is lower than 15%. It does not allow a good screening step. The molecular structures present in this area are the acyclic structures or structures with only little few sequences of cyclic and acylic sets. For these structures, the $RG^{(cycle)}$ concept is not efficient. One could propose in such a case to use another type of reduction. A proposition in that sense is done in the final part of this paper.

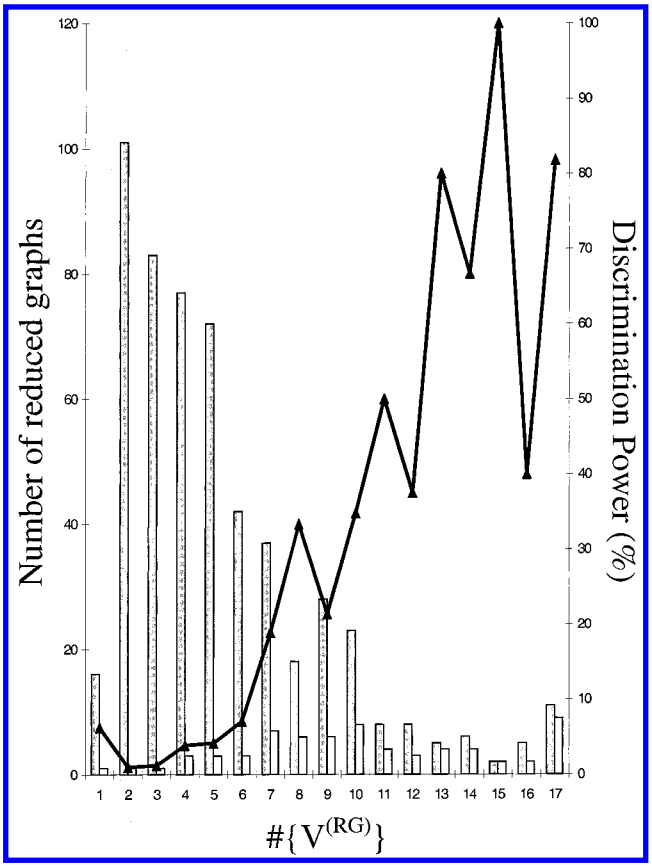

Figure 15. Histogram showing the action of the *RG* based on the carbon/heteroatom distinction over a test population of 540 molecular structures. The abscissa values are the number of vertices per RG, the ordinate values, the number of RG. Gray columns represent the total number of graphs, white columns, only the different graphs. Black triangles symbolize the value of the discrimination power.

6. DISCUSSION AND CONCLUSION

Our aim was to obtain a good molecular structure descriptor for the efficient and reliable screening step of structures and substructures present in standard organic reaction databases. As cyclic structures are widely present in organic chemistry, an obvious first idea is to work with graphs based on the cyclic information contained in the molecular structures. But as we did not find any ring set definition allowing to describe all the simple rings of the planar projection of each molecular structure, we developed a new set of rings as well as the corresponding reduced graph. The proposed set of rings can be obtained with a recurrent method. We start with the computation of the first iteration set, the Smallest Cycles at Edges, $SCE^{(1)}$, by searching the smallest cycles for each edge of the cyclic subgraph. In a second stage, the internal rings are added to the set of rings with the other iterations. With these additional rings, the description of all the planar simple-cycles is now possible. The obtained ring set is called the Set of Smallest Cycles at Edges. It is a balance between the loss of information (by keeping only certain cycles) and the ease of substructure retrieval.

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.

To analyze the power of discrimination of our $RG^{(cycle)}$, the concept was tested over a test database of 540 molecular structures. According to the results of the histogram analysis, 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 Figure 13), the $RG^{(cycle)}$ concept is clearly not efficient. Another kind of reduction should then be used. As a test, we decided to apply a reduction based on the carbon/heteroatom distinction to our 540 organic structures test database (Figure 15). Again, the total histogram can be divided in three regions according to the discrimination power. But, the most important interest in the use of a second type of RG is that the structures that appeared in the A, B, and C regions of Figure 13 may now appear in other regions. Indeed, as shown in Figure 16, a structure that showed up in the $RG^{(cycle)}$ within area C, i.e., the least discriminating one, may now appear in area B, with a higher discrimination power, when another criterion of reduction, the carbon/heteroatom criterion, is used.

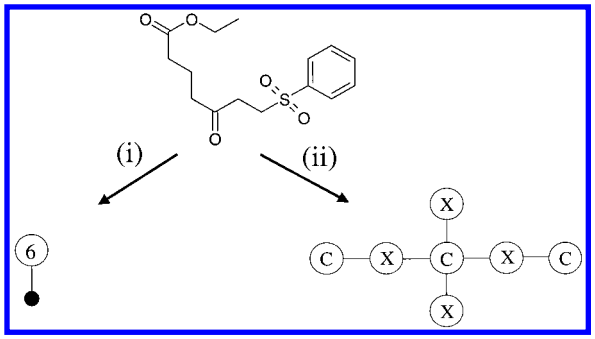

Figure 16. Comparaison between the $RG^{(cycle)}$ (i) and the *RG* based on the carbon/heteroatom distinction (ii) for a particular molecular structure present in the test database. The label of the open circles of the $RG^{(cycle)}$ correspond to the size of the rings reduced in each of the *GSCE* vertex. The black circles of the $RG^{(cycle)}$ represent the vertex of the GAS. The letter in the open circles of the *RG* based on the carbon/heteroatom symbolize the vertex containing the carbon atom (C label) and the vertex containing heteroatoms (X label).

In conclusion, the histograms allow us 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 processed. Indeed, the discrimination power of a type of reduced graph is dependent on the database that is used. At each introduction or deletion of a molecular structure, 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. The use of reduced graphs as screening descriptors is thus a necessary step to allow a fast and efficient search for structure or substructure searching.

ACKNOWLEDGMENT

L.D. thanks Dr. G. M. Downs for his valuable help in commenting this paper before submission, Prof. A. Krief for useful discussions, the CSA trust for its bursary to participate to the Fifth International Conference on Chemical Structures, and the "Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture" (FRIA) for his PhD fellowship.

REFERENCES AND NOTES

(1) Corey, E. J.; Wipke W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166*, 179−192.

(2) Barth, A. Status and Future Development of Reaction Databases and Online Retrieval Systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 384−393.
(3) Blake, J. E.; Dana, R. C. CASREACT: More than a Million Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 394−399.
(4) Parlow, A.; Weiske, C. ChemInform: An Integrated Information System on Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 400−402.
(5) Jochum, C. The Beilstein Information System is not a Reaction Database, or is it? *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 71−73.
(6) Wiggins, G. Caught in a CrossFire: Academic Libraries and Beilstein. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 764−769.
(7) 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.* **1993**, *33*, 539−541.
(8) Qian, C.; Fisanick W.; Hartzler D. E.; Chapman S. W. Enhanced Algorithm for Finding the Smallest Set of Smallest Rings. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 105−110.
(9) Downs, G. M. Ring Perception. *The Encyclopedia of Computational Chemistry*; von Ragué Schleyer, P., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley and Sons: Chichester, 1998 Vol. 4, pp 2509−2515.
(10) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Theoretical Aspects of Rings Perception and Development of the Extended Set of Smallest Rings Concept. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 187−206.
(11) Fujita, S. Logical Perception of Ring Opening, Ring Closure, and Rearrangement Reactions Based on Imaginary Transition Structures. Selection of Essential Set of Essential Rings (ESER). *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 1−9.
(12) 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.* **1987**, *27*, 126−137.

CI000401Y