

## Representation of the Molecular Topology of Cyclical Structures by Means of Cycle Graphs. 3. Hierarchical Model of Screening of Chemical Databases

Irene Luque Ruiz,\* Gonzalo Cerruela García, and Miguel Ángel Gómez-Nieto

Department of Computing and Numerical Analysis, University of Córdoba,  
Campus Universitario de Rabanales, Albert Einstein Building, E-14071 Córdoba, Spain

Received March 30, 2004

The increase in the size and complexity of chemical databases necessitates the proposal and development of efficient methods of classification and recovery of information, which supposes proposal of a model of classification of database records and the use of a compatible model of screening for inspection of clusters and recovery of the molecules that satisfy the search criterion. The cycle graphs model based on consideration of all the cycles and chains (and equivalent cycles and chains) present in the molecular structure has been proven appropriate for classification of chemical databases, giving rise to a generation of different classification levels depending on the structural elements (cycles and chains) that are considered. In this paper we propose a screening model, compatible with the cycle graphs model, based on a hierarchy of levels of abstraction. The set of molecules that satisfies a screening model (or selection criterion) diminishes as we advance in the hierarchy of levels of the model, which allows filtering of records and, therefore, an increase in the efficiency of the screening process. In the following work of this series we describe and validate the screening tool developed.

### 1. INTRODUCTION

In a molecular graph  $G$  the nodes represent the atoms and the edges represent the connections (bonds) among the atoms in the molecule. Molecular graphs can be represented in many ways, as adjacency matrix, connections matrix, distance matrix, etc. On occasion it is convenient to use homomorphic or isomorphic representations of the molecular graph. These representations  $G'$  allow us to represent the structures or topologies of the molecules without a significant loss in the information to be studied.<sup>1</sup>

The cycle graphs (CG) and cycle and chain graphs (CCG)<sup>2,3</sup> (molecular graphs based on the cyclicity of the structure) represent a clear example. These graphs have demonstrated their utility for extraction of topological invariants contributing measures of complexity, cyclicity, and symmetry of the molecular structures and in the clustering of chemical databases.

Making use of the CG and CCG graphs the chemical databases are classified under a hierarchical paradigm that considers, at a different level of abstraction, the structural elements (cycles and chains) present in the molecules. In this way, efficient access to information in classified databases under this hierarchical model requires an appropriate model of screening.

By screening a known process, based on a search criterion, a set  $S$  of records satisfying this criterion is recovered from the database with the objective, later on, to analyze this set of records and/or refine this set by means of matching processes (of any type) with a pattern problem.<sup>4–6</sup>

Evidently as the search criterion is more general, the cardinality of the set  $S$  is greater and vice versa. A very

general criterion will give place to a set  $S$  with many elements which will be translated to a computational cost arising in the later matching process. However, a very restrictive search criterion will cause a set  $S_1$  with very few elements, producing an interesting set  $S_2$  of elements ( $S_1 + S_2 \subset S$ ) which are not recovered.

Thus, it is convenient that in the screening processes we have tools/models that allow to refine the search approach under the same paradigm. This way we will be able to obtain an initial set  $S$  which we will be able to refine (to decrease) by means of the specialization (details) of the search approach, in a process of steps, until reaching an appropriate cardinality of the set  $S_1$  for the later matching process.

In this paper an abstract model of screening is described that considers the hierarchical paradigm of classification of databases based on the cycle and chain graphs.<sup>2,3</sup> The screening model is based on consideration of the structural entities—cycles and acyclic chains—present in the molecular graphs and in representation, at different levels of abstraction, of the molecular shape using these structural entities by means of an easy and simple graphic notation.

The molecular shape is quite an abstract concept unless we consider the physical measurements of the distances and angles through which the atoms in the molecule are connected. Since this information is only known for those well-studied compounds, the shape of the molecules is usually defined as an abstract measure obtained from topological invariants derived from the molecular graph that represents the molecular structure of a chemical compound.<sup>7–9</sup>

However, if we only take into account the structural entities —cycles and acyclic chains—present in the molecular graph, we can build an abstract representation that represents the topological shape of the molecule and use this repre-

\* Corresponding author phone: +34-957-21-2082; fax: +34-957-21-8630; e-mail: ma11urui@uco.es.

sensation as a selection approach in the screening process. In this paper, this screening model is described as well as the classification model based on the cycle and chain graphs which considers these structural entities.

The manuscript has been organized in the following way: Initially we give a brief review of the cycle and chain graph models showing the hierarchy of the models. Section 3 describes the screening model, detailing each of the levels of abstraction on which it is based and their correspondence with the classification models. Last, we discuss applications of the proposed model, leaving the description and test of the tool developed in Java language implementing the screening model described here for the following article of this series.

## 2. BACKGROUND

In a previous article<sup>3</sup> we proposed the use of a homomorphic representation of molecular graph  $G$  for representation of the topological structure of the molecules, which can be used conveniently for the extraction of topological descriptors and their application in the processes of clustering of chemical databases. These homomorphic representations of the molecular graph are called, in a general way, cycle graphs (CG) and cycle and chain graphs (CCG) (see Abbreviations section for details of the abbreviations used).

The CG and CCG graphs represent, by means of nodes and edges, basic structural elements of the molecular topology. Depending on the type of cycle graphs, these structural elements are<sup>3</sup>

- (1) only the cycles (CG model);
- (2) only the equivalent cycles (ECG model);
- (3) the cycles and the chains without considering their size (CCG<sub>1</sub> model);
- (4) the equivalent cycles and chains without considering the size of the chains (ECCG<sub>1</sub> model);
- (5) the cycles and the chains considering the size of the chains (CCG<sub>x</sub> model);
- (6) the equivalent cycles and chains considering the size of the chains (ECCG<sub>x</sub> model).

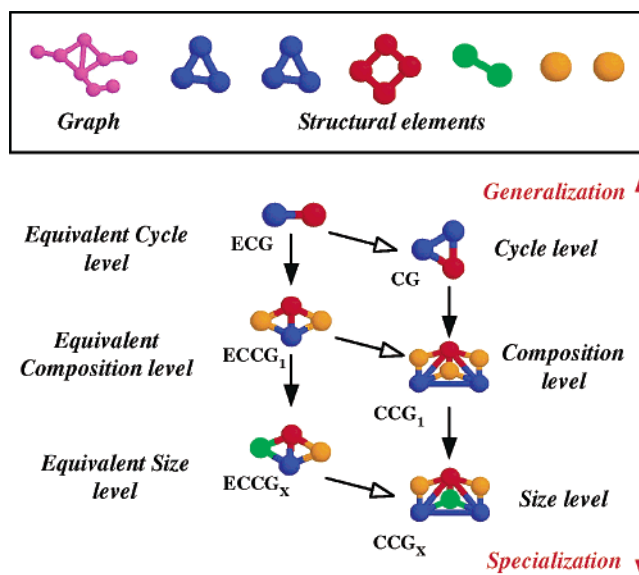
Construction of these graphs is carried out as follows:

Given a cyclical structure represented by a molecular graph  $G$ , all present cycles in the graph can be extracted efficiently using the algorithms proposed by the authors.<sup>10</sup> Knowing the set of cycles  $C$  present in a molecular graph  $G$ , a homomorphic graph of cycles GC can be built with the following characteristics.

- (1) The GC graph is a undirected, weighted, and colored graph.
- (2) The GC graph has the same number of nodes as elements are present in the  $C$  set. Each node is identified by the size of the cycle (number of nodes in the  $G$  graph) represented by the node.
- (3) The edges of the GC graph are labeled, and they represent the relationships among cycles of the  $C$  set. Given two nodes  $i, j \in GC$ , which represent the cycles  $C_i$  and  $C_j$  existing in  $G$ , the edge that relates both nodes is labeled with the number of common nodes to the cycles  $C_i$  and  $C_j$ , that is,  $C_i \cap C_j$ .

The graph of equivalent cycles ECG is built,<sup>2,3</sup> a homomorphic graph to the GC graph, in the following way.

- (1) The ECG graph is a undirected, weighted, and colored graph.



**Figure 1.** Example of molecular structure and its corresponding structural elements (cycles and chains). The corresponding graphs GC, ECG, CCG<sub>1</sub>, ECCG<sub>1</sub>, CCG<sub>x</sub>, and ECCG<sub>x</sub> and the related representation level in the screening model are shown.

(2) The ECG graph is composed of as many nodes as equivalent class of cycles present in CG graph.

(3) The edges of the ECG graph are labeled and represent the relationships—number of common nodes—among the equivalent class of cycles in CG.



Once all the nodes that are part of the cycles in the  $G$  graph are known, the chains are extracted, that is, the set of nodes of  $G$  that are connected to each other and are not part of a cycle. Each independent chain is represented as a node in the CCG graph, which will only be connected (related) with other nodes corresponding to cycles.

The ECCG graph is built from a CCG graph, obtaining the classes of equivalence existing in CCG (taking into account that the nodes corresponding to cycles are colored in different ways than the nodes corresponding to chains) and carrying out the same process aforementioned for the construction of the ECG graph.






Figure 1 shows the hierarchy of abstraction proposed by our paradigm. The CCG<sub>x</sub> model considers the graph structural elements with a higher level of detail. This model considers the size of the cycles and the chains, resulting in greater specialization. Consideration of the equivalent classes of cycles and chains contributes a higher generalization (less detail), giving rise to the ECCG<sub>x</sub> model. The ECCG<sub>x</sub> model only considers the cycles and chains that present the same behavior in the  $G$  graph.

When the size of chains is not considered, we advance in the generalization process. In the CCG<sub>1</sub> model the chains of  $G$  graph with different sizes are represented in the CCG<sub>1</sub> graph with nodes with equal labels. Thus, the CCG<sub>1</sub> model, although taking into account the presence of all the chains, supposes that the size does not influence the behavior of the information represented by the  $G$  graph. A step further in the generalization process is contributed by the ECCG<sub>1</sub> model, which only considers those chains and cycles that have a different behavior.

The ECG model is the most general of all. This model only considers the equivalent classes of cycles present in

Bars	Lines
	
Representing cycles	Representing acyclic chains

**Figure 2.** Graphic elements used for the representation of the topological or structural shape (TS).

Acyclic chains	Cycles	Cycles and chains
 $\text{CH}_3\text{-CH}_2\text{-CH}_3$	 	  $\text{-CH}_2\text{-CH}_3$

**Figure 3.** Example of the use of graphic elements for representation of TS for some molecules. Chains are represented by lines, and cycles are represented by bars.

the G graph and is specialized in the CG model. Neither model considers the present chains in G, supposing that for the study and analysis of the information represented by the G graph these elements are not outstanding.

As we can appreciate, the different representation models generate a hierarchy of abstraction, from a very general model (ECG) to a specialized model (CCG<sub>X</sub>) which considers the structural characteristics of the G graph and, therefore, all the characteristics of the entities and relationships corresponding to the domain of the problem and represented by the G graph.

### 3. ABSTRACT MODEL FOR REPRESENTATION OF THE TOPOLOGY OF MOLECULAR STRUCTURES

The classification model based on the cycle and chain graphs makes use of structural elements of the molecules (cycles and chains), which supposes that the recovery processes should keep in mind these same structural elements and under a compatible paradigm. In this paper we propose use of the topological shape of the molecules as a search approach in the screening processes. To do this it is necessary to define *what is considered as a topological shape*.

We define the structural shape or topological shape (TS) as a graphic representation based on default basic structures that represent the existence of cycles and chains present in the molecular structure. As observed in Figure 2, this graphic representation uses two basic structures.







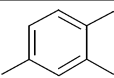
Using these basic structures of representation (bars and lines) we propose a hierarchical model based on the abstraction principles<sup>11</sup> by means of which the topological shape of the molecules can be represented at different detail levels.

The representation will be carried out as follows (Figure 3):

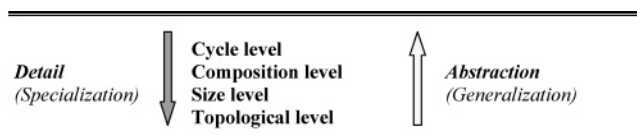
- (1) The acyclic chains will be represented graphically by a horizontal line (line).
- (2) The cycles will be represented by a vertical line (a bar).
- (3) If acyclic chains and cycles exist in the molecule, they are represented by the combination of lines and bars.

In the proposed model the lines representing chains will always be placed on the right side of the bars, and their position (in height) with regard to the cycles has no meaning. In Figure 4 some correct and incorrect representations are shown for a molecule example.

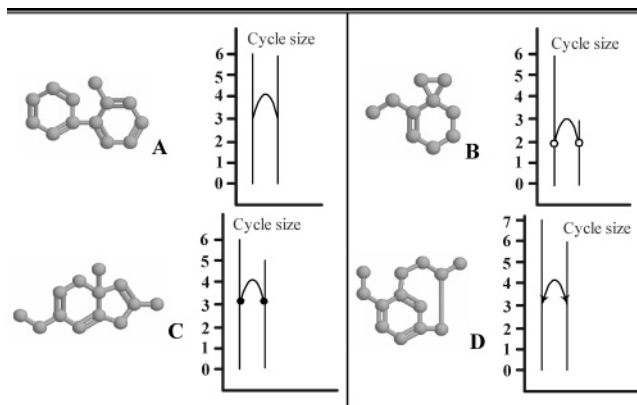
As shown in Figure 5, the proposed model is based on the representation of the topological shape in four possible levels of abstraction.<sup>11</sup> As we increase the detail (higher specialization) the corresponding topological shape (TS)

Correct	Incorrect	Molecule
  	  	

**Figure 4.** Correct and incorrect use of the graphic elements for representation of TS for a molecule example.



**Figure 5.** Hierarchy of abstraction for the different levels of representation of TS



**Figure 6.** Some TS for the cycle level with an example of molecules which satisfy the corresponding TS: (A) two cycles without common nodes (an arch is used), (B) two cycles with a common node (an arch with empty cycles is used), (C) two cycles with two common nodes (an arch with filled circles is used), and (D) two cycles with three common nodes (an arch ending in arrow is used, the as same when more than three common nodes are present). At this level chains are not considered.

corresponds to a smaller number of possible molecular structures and vice versa; as the abstraction increases (higher generalization), the TS corresponds to a greater number of possible molecular structures.

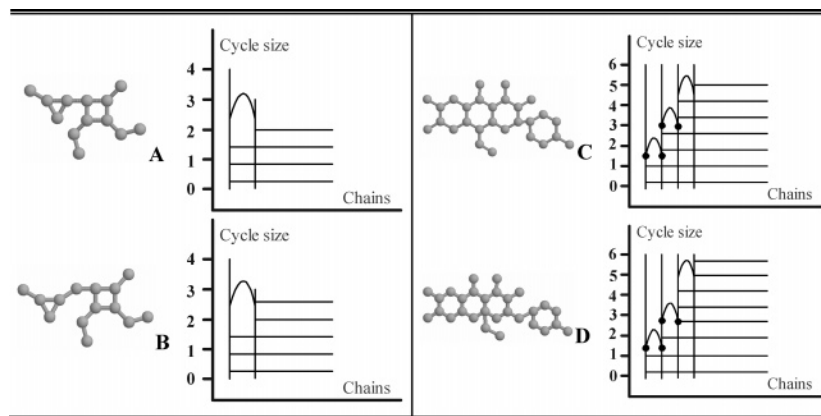
It can be deduced from the proposed model that in a screening process based on the definition of search pattern definition by the TS, the set S of recovered molecules that satisfy the search pattern will be smaller as the utilized level of abstraction decreases and vice versa, that is, a search pattern represented by means of a TS based on the size level will produce a smaller cardinality of the set S than a search pattern based on the cycle level.

**3.1. Cycle Level.** The cycle level is the highest level in the hierarchy of the abstraction in which fewer details of the molecular structure are represented and a set S will be obtained with a higher cardinality when carrying out the screening process for a given pattern.

At this level only the existence of present cycles in the molecules by means of the graphic elements described in Figure 2 is represented. The representation is carried out in a one-dimensional frame in which the cycles and the relationships among cycles in the molecule are represented.

In this frame the size of each cycle is shown (Y axis), and to carry out the representation of the relationships among the cycles, the common atoms among each couple of cycles are considered. Thus, as Figure 6 shows, the following cases are considered:





**Figure 7.** Some TS for the composition level with an example of molecules which satisfy the corresponding TS.

(1) Cycles that do not have any nodes in common. In this case the cycles can belong to independent components<sup>2</sup> or are connected by a bond or an acyclic chain. This relationship is represented by an arch joining the cycles.

(2) Cycles have one node in common. This relationship is represented by an arch ending in empty circles that join the cycles.

(3) Cycles have two nodes in common. This relationship is represented by an arch ending in filled circles that join the cycles.

(4) Cycles have more than two nodes in common. This relationship is represented by an arch ending in an arrow tip that joins the cycles.

As can be observed, at this level the presence of chains is not represented but rather only shows the cycles and relationship among the cycles.

**3.2. Composition Level.** The composition level allows representation of all the structural elements that can be present in a molecule (cycles and chains) in the TS. It is an advance in the detail or specification of the topological shape in which the information of the cyclicity of the molecules and the relationships among the cycles is represented (as in the previous level), and also, the existence of acyclic chains and their relationships with the cycles is represented. However, at this level any detail is not due to the size of the chains.

This representation level is carried out in a two-dimensional frame in which

(1) the Y-axis is used to represent the size of the cycles, the same as in the previous level, and

(2) the X-axis is used to represent the chains. It is a literal or informative axis.

Although at this level information of the size of the acyclic chains is not represented, the relationship that these chains maintain with the cycles is. Thus, a chain can be related with a cycle in two ways:

(1) attached to the bar that represents a cycle, describing an acyclic chain that is connected to some atom of the cycle that is not common to another cycle;

(2) attached to the arch that relates two cycles, describing an acyclic chain that is connected to some atom common to two cycles.

As shown in Figure 7, at this level besides indicating the size of the cycles in the TS, the existing relationship between the cycles and the detail of the relationship type between the cycles and the chains is represented. However, details of the size of the chains are not included.

Figure 7 shows the different relationships that the chains can have with the cycles. In the graph in part A the cycles are related directly by a connection, being represented by the arch that joins them, while in the graph in part B the cycles are related through an atom, which is represented by a chain connected to the arch that relates both cycles.

On the other hand the graph in Figure 7C shows a connected system of six node rings: (a) the first three cycles (from left to right) maintain two nodes in common, (b) the third and fourth cycle are related through a connection, and (c) common chains to two cycles do not exist. However, in the graph in part D (a) the third and fourth cycle are related through an atom of carbon, which is represented by a chain connected to the arch that relates both cycles in the TS, and (b) the second and third cycle (from left to right) present a chain joined to a common node, which again is represented in the TS by a line joined to the arch that relates both cycles.

**3.3. Size Level.** At this level we represent (in the TS with a full level of detail) the structural entities of the molecules, that is, information of the cycles and chains with details of their relationships and size, contributing to complete detail of the cyclic and acyclic substructures of the topology of the molecule.

The representation is close at the composition level, but in this case the X-axis is a numeric axis in charge of representing the size of the acyclic chains present. In Figure 8 some examples are shown for the size level.

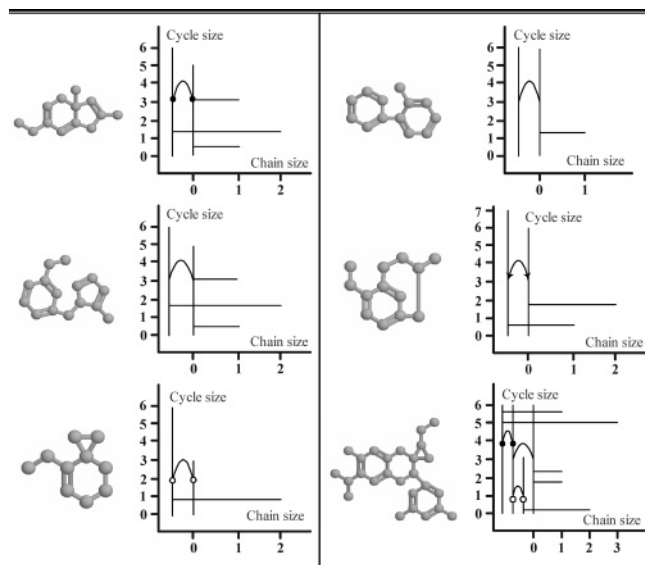
**3.4. Correspondence Among Representation Levels.** As we can observe, the proposed model for representation of the TS is a hierarchical model from the point of view of abstraction (see Figure 1). A TS in a cycle level is more general than that in a composition level and this in turn is more general than that in a size level, which implies that:

(1) with  $TS_X$  a TS in the cycle level,  $TS_C$  a representation in the composition level, and  $TS_S$  a representation in the size level, the set of molecular structures  $M$  (molecules) that satisfies these representations increases in the order  $TS_X > TS_C > TS_S$ .

(2) Given a  $TS_X$ , there are many  $TS_C$  that satisfy this  $TS_X$ , and given a  $TS_C$ , there are many  $TS_S$  that satisfy this  $TS_C$ .

And vice versa, a TS in a size level is more specialized than in a composition level, and this in turn is more specialized than in a cycle level, which implies that:

(3) given a  $TS_S$  there is one and only one  $TS_C$  that satisfies this  $TS_S$ , and given a  $TS_C$  there is one and only one  $TS_X$  that satisfies this  $TS_C$ .



**Figure 8.** Some TS for the size level with an example of molecules which satisfy the corresponding TS.

**3.5. Correspondence between the Levels of Representation of the TS and Cycle Graph Models.** The pattern of representation of the molecular graphs based on the cycle graphs is also a hierarchical model from the point of view of abstraction.<sup>11</sup> As shown in Figure 1, as we consider more characteristics in the structural elements of the molecular topology (cycles and chains), the specialization of the representation model increases. Thus, a  $CCG_X$  graph is more specialized than a  $CCG_1$  graph, and this in turn is more specialized than a CG graph.

Therefore, similar to the proposed model TS, a CG graph represents many  $CCG_1$  graphs and in turn many  $CCG_X$  graphs, while a  $CCG_X$  graph represents a unique  $CCG_1$  graph, and this in turn a unique CG graph.

As Figure 1 shows, a clear correspondence exists among the different levels of the TS model and the cycle and chain graphs in the way:

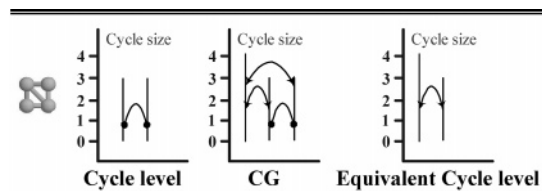
(1) the cycle level represents CG graphs, graphs whose nodes represent cycles of the molecule and whose edges represent the common nodes among the cycles.<sup>2,3</sup>

(2) The composition level represents  $CCG_1$  graphs, graphs whose nodes represent cycles and chains (without considering the size) of the molecule and whose edges represent the common nodes among the cycles and between cycles and chains.<sup>3</sup>

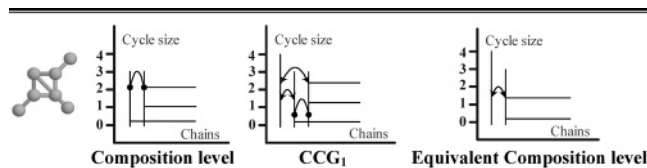
(3) The size level represents  $CCG_X$  graphs, graphs whose nodes represent cycles and chains (take into account the size) of the molecule and whose edges represent the common nodes among the cycles and between cycles and chains.<sup>3</sup>

As Figure 1 shows, the cycle graphs model considers three other homomorphic graphs to the molecular graph that are on an intermediate abstraction level. The nodes of the ECG,  $ECCG_1$ , and  $ECCG_X$  graphs represent the equivalent classes (cycles and chains) present in the CG,  $CCG_1$ , and  $CCG_X$  model, respectively, which causes an increase in the generalization at the representation level with regard to the corresponding graph (Figure 1).

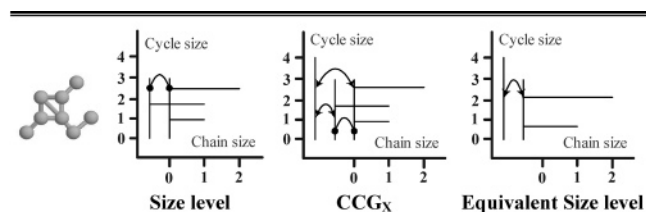
Our representation model based on the structural or topological shape allows immediately obtaining these three intermediate abstraction levels by obtaining all the present



**Figure 9.** TS at the Cycle level with a molecule example which satisfies the representation. Central columns shows the corresponding representation in the CG model, and right column shows the generated representation at the Equivalent Cycle level.



**Figure 10.** TS at the Composition level with a molecule example which satisfies the representation. Central columns shows the corresponding representation in the  $CCG_1$  model, and right column shows the generated representation at the Equivalent Composition level.



**Figure 11.** TS at the Size level with a molecule example which satisfies the representation. Central columns shows the corresponding representation in the  $CCG_X$  model, and right column shows the generated representation at the Equivalent Size level.

cycles in the graph representing the TS. Thus, it is possible to obtain the ECG level starting from CG,  $ECCG_1$  starting from  $CCG_1$ , and  $ECCG_X$  starting from  $CCG_X$  as shown in Figures 9–11 for the graphs of the Figure 1.

Starting from a TS model, in the cycle level containing two cycles of three nodes connected by an arch with the filled circles (the cycles maintain two nodes in common), a new cycle formed by four nodes can be derived which maintains three nodes in common with each of the cycles of three nodes (see Figure 1). Therefore, under the cycles and chains model,<sup>3</sup> the TS model of the cycle level represents all the molecules where the structure shown in the central column of Figure 9 is present.

The existence of two classes of equivalent cycles can be appreciated: a class corresponding to the cycle of four nodes and another class corresponding to the two cycles of three nodes. This information can be extracted in a very simple way<sup>3,12</sup> and, therefore, automatically generates the equivalent cycle level shown in Figure 9.

Figure 10 shows an example for a TS based on the composition level. This TS represents any molecule in which two cycles of three nodes sharing two common nodes and three chains of any size joined to the cycles are present.

Similar to the previous example, consideration under the cycle and chain model of a new cycle of four nodes gives us the  $CCG_1$  representation (central column in Figure 10) from which the existence of two equivalent cycles can be extracted (the cycles of three nodes) and two equivalent

Topological Index	A	B	C	D	E	F	G	H	I	J	K
<sup>1</sup> $\chi$ Randic	4.43	4.43	4.43	4.43	3.00	3.00	3.00	1.41	1.41	1.41	1.41
Wiener	94.00	94.00	94.00	94.00	27.00	27.00	27.00	4.00	4.00	4.00	4.00
Balaban (J)	2.08	2.08	2.08	2.08	2.00	2.00	2.00	1.63	1.63	1.63	1.63
Kier & Hall	4.50	4.50	4.36	4.36	3.00	3.00	2.95	1.50	1.50	1.41	1.41
Pogliani	18.00	18.00	19.50	19.50	12.00	12.00	12.50	6.00	6.00	7.00	7.00
Kier shape	7.11	6.09	5.98	6.23	4.17	3.41	3.34	3.00	2.74	2.70	2.96
Zagreb	144.00	88.00	71.00	85.00	96.00	54.00	49.00	48.00	34.00	22.00	36.00
Balaban (J <sub>Z</sub> )	2.08	3.02	3.19	2.74	2.00	3.00	3.19	1.63	2.19	2.62	1.87
Wiener (W <sub>V</sub> )	3.98	5.20	5.24	4.81	3.74	5.18	5.14	2.58	3.41	3.71	2.72

Figure 12. Values of different topological invariants for four molecules and their basic elements (cycles and chains).

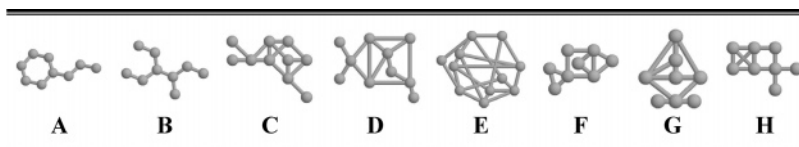


Figure 13. Molecular graphs obtained with the developed algorithm<sup>17</sup> for values of the Wiener index = 94 (A–E) and 51 (F–G).

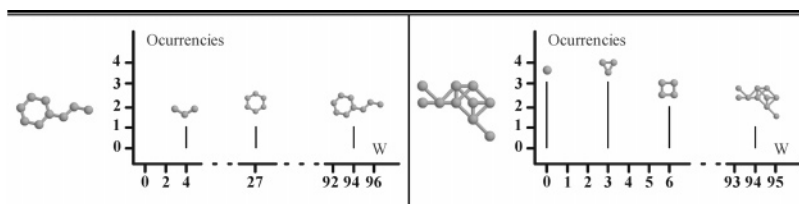


Figure 14. Topological level corresponding to the molecular graphs A and C of Figure 13. Molecules are fragmented, and their basic elements (cycles and chains) are also considered.

chains (the chains not joined to the common node to the cycles of three nodes). This information allows generating the equivalent composition level as shown in Figure 10.

Last, Figure 11 shows an example of a TS of the size level for which the corresponding CCG<sub>X</sub> can be derived and, once the classes of equivalent cycles and chains are obtained, the corresponding equivalent size level can also be obtained.

**3.6. Topological Level.** The last representation level that considers the TS model is the topological level. It is the most specialized level from the point of view of the abstraction in which, with maximum detail, the structural elements (cycles and chains) considered in the TS with a model are represented.

While in the previous level (size level) a TS model represents a wide set of molecules which contain the structural elements (and their relationships) described in the model, at the topological level the presence of the structural elements of the set of molecules that satisfies a certain model will depend on the topological invariant used at this representation level, as we will see later on.

The representation at this level has the following characteristics:

- (1) The structural elements that are represented are the cycles and chains (the same as in the previous described levels).
- (2) The representation is carried out in a two-dimensional frame. The X-axis represents each of the different types of structural elements (cycles and chains) on a scale determined by the topological invariant selected.
- (3) The Y-axis represents the occurrences (number of cycles and chains) that satisfy a certain value of the topological invariant selected.

The topological invariant used in the X-axis can be any, determining the characteristics of the invariant the detail or specialization of the representation. This characteristic is observed in Figure 12, which shows the values of different topological invariants for four molecules (A–D) which have the same structural entities (a cycle of six nodes and a chain of three nodes). These molecules present the same TS for the previous three levels described (cycles, composition, and size), although the molecules are different. Also, Figure 12 shows the values of selected topological invariants<sup>7,8,12–16</sup> for each of the topological entities (six nodes cycle and three nodes chain).

As observed, all molecules (A–D), all cycles (E–G), and all chains (H–K) have the same value of the Wiener, Balaban distance connectivity (J) and Randic connectivity indexes which does not allow differentiation between either of the molecules or the different types of structural elements (cycles and chains). This situation would also take place with other topological invariants that do not take into account the color of the nodes and/or the edges in the molecular graph (any topological descriptor calculated from the adjacency matrix, distances matrix, etc.).

Other topological invariants such as the Kier and Hall or Pogliani index consider the color of the nodes, giving different values for those graphs in which heteroatoms are present. Other descriptors such as Kier shape, Zagreb, Balaban index (J<sub>Z</sub>) using Z-weighted distance matrix (Barysz matrix), or valence Wiener index (W<sub>V</sub>) are able to differentiate all the structures shown in Figure 12.

Thus, selection of the topological descriptor will determine the discrimination power at the topological level. As it increases the capacity of discrimination of the topological

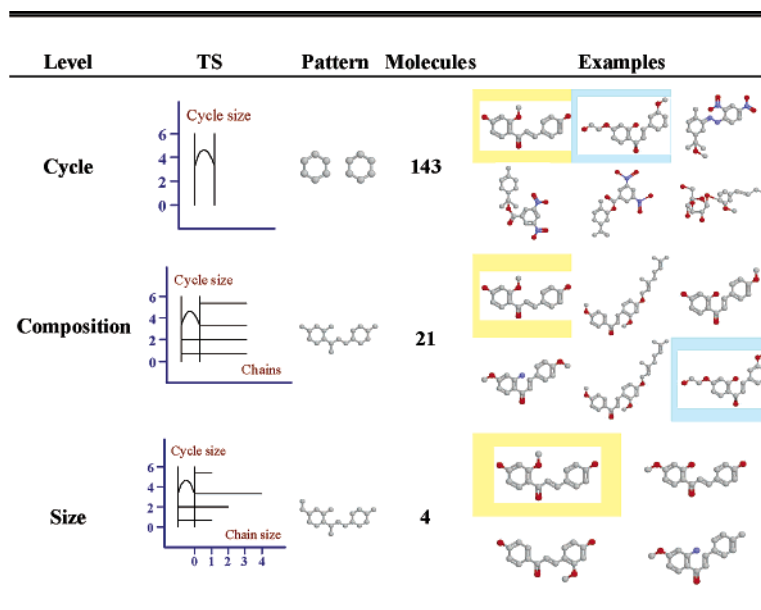


Figure 15. Example of validation of the screening model considering the main abstraction level (cycle, composition, and size).

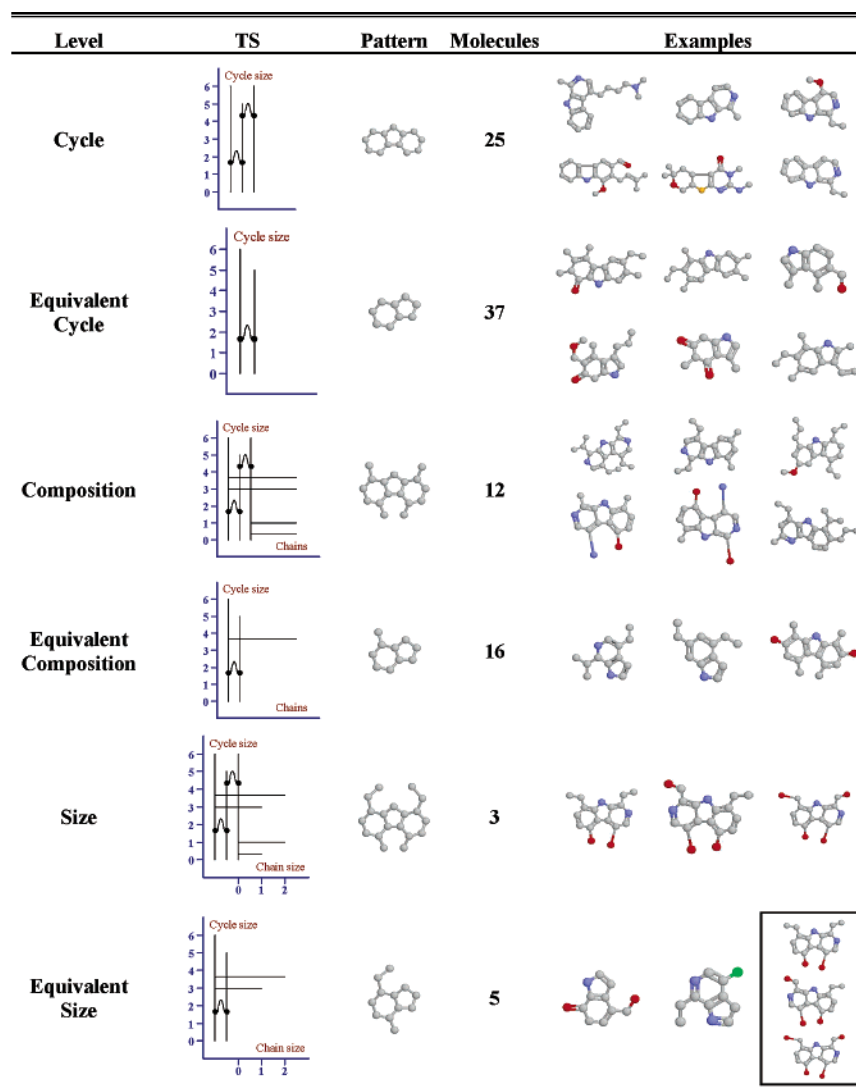


Figure 16. Behavior of the different TS in the screening process. The hierarchy of abstraction for the levels and equivalent levels of representation determines the set of molecules recovered.

descriptor utilized, the number of molecules that satisfies the TS at the topological level will be smaller and vice versa.

Evidently development of a computational tool to implement the topological level is complex. This tool should be



able to generate, with an acceptable computational cost, a set of molecular structures that satisfies a given value of the selected invariant.

The computational complexity of this process depends mainly on two factors: (a) the topological invariant selected and (b) the value of this invariant. To diminish the computational cost of this process, it is necessary to develop efficient heuristics which diminish the combinatorial explosion that bears obtaining all those molecular structures that satisfy a certain value of the topological invariant selected, as the authors proposed.<sup>17</sup>

Figure 13 shows the molecular graphs obtained with the developed algorithm<sup>17</sup> for values of the Wiener index equal to 94 (A–E) and 51 (F–H). Evidently the number of molecules that satisfy a certain value of the topological invariant depends on the parameters used in the calculation of the invariant (heteroatoms, cycles number and size of cycles, number of double and triple bonds, etc.) and the value of the selected invariant. To clarify the example shown in Figure 13, only simple bonds, any cycle number, and nonpresence of heteroatoms have been considered.

In Figure 14 the topological level corresponding to molecular graphs A and C of Figure 13 are shown. As can be observed, the A molecule has two fragments: (a) a chain of three nodes ( $W = 4$ ) and (b) a cycle of six nodes ( $W = 27$ ). On the other hand, the C molecule has three different fragments: (a) three chains of one node ( $W = 0$ ), (b) three cycles of three nodes ( $W = 3$ ), and (c) two cycles of four nodes ( $W = 6$ ). Extraction of the topological entities from a molecular graph is carried out by extraction of all the chains and the set of cycles SSSR (Smallest Set of Smallest Rings)<sup>18</sup> present in the molecular graph.

**3.7. Validation of the Proposed Model.** Validation of the model has been carried out on a database of 2741 molecules.<sup>19</sup> This database has been previously classified under the different models based on the information of the cycles and chains (CG, CCG<sub>1</sub>, CCG<sub>X</sub>) and equivalent cycles and chains (ECG, ECCG<sub>1</sub>, ECCG<sub>X</sub>).<sup>3</sup>

The process consists of the selection of a pattern object of the search and construction of the different representations (TS) for each of the abstraction levels considered by the model (cycles, composition, and size), which are used as search criterion in the database.

Figure 15 shows an example of the different TS for the selected molecule, the pattern search corresponding to each level, the number of recovered molecules, and some examples of these molecules. As we can appreciate, the number of recovered molecules increases with the use of TS based on higher abstraction. Thus, the size level is more selective than the composition level and this, in turn, is more selective than the cycle level. As observed in Figure 15, the molecules recovered in the size level (an example has been colored in yellow) are recovered in the composition and cycle levels. In turn, other molecules recovered in the composition level (an example has been colored in blue), not recovered in the size level, are also recovered in the cycle level.

Another validation example considering the intermediate abstraction levels based on the equivalent cycles and chains is shown in Figure 16. As this figure shows, the levels based on consideration of equivalent cycles and chains contribute an intermediate refinement level, generating a new refinement among the size, composition, and cycle levels, as shown for

the equivalent size level in which five molecules are recovered, three of them (represented in a frame) are also recovered in the size level, behavior described in a previous paper.<sup>2</sup>

#### 4. DISCUSSION AND REMARKS

Throughout the manuscript an oriented graphic model for representation of the structural or topological entities present in the molecular graphs has been described. This model is based on consideration of the cycles and chains of the molecular graph and representation of these entities and their relationships of simple graphic symbols: bars, lines, and arches.

On the basis of the fundamental principles of the abstraction, the proposed model allows representation of the structural elements of the molecular graphs at different detail levels under a hierarchical paradigm, advancing from a very general level (cycle level)—where only the cycles and their relationships present in the molecular graph are considered—to a detailed level (size level)—where all topological entities of the molecular graph (cycles, chains, size, and relationships) are considered.

The first three representation levels (cycle, composition, and size levels) can be generalized to the corresponding intermediate levels of abstraction (equivalent cycle, composition, and size levels) by consideration of the equivalent classes of cycles and chains present at the respective levels.

Therefore, the six representation levels (cycle, composition, and size and their corresponding levels based on the equivalent classes of cycles and chains) considered in the proposed model present a full integration and compatibility with the classification model proposed by the authors,<sup>3</sup> equally based on the consideration of the cycles and chains present in the molecular graph. This compatibility allows the screening model proposed in this article to be used for development of a screening tool for the search and efficient recovery in chemical databases.

The screening model proposed is characterized by the great interconnection among the different representation levels, which means that a representation at a detail level (size level) can be directly translated at the most general level (composition level)—and one representation at the composition level to the corresponding representation in the cycle level—due to the hierarchical paradigm on which the model is based.

In addition, the screening model proposed advances the detail level a step further with regard to the classification model based on the cycle and chain graphs.<sup>3</sup> The topological level allows consideration of different topological invariants for the selection of molecular structures/substructures from which a recovery process can be carried out to be used for representation of a screening approach at a more general level of detail (size, composition, or cycle).

Selection of the topological invariant at the topological level will determine the specialization or detail of the screening approach using this level. Thus, an invariant that does not consider the color of the nodes, nor of the connections type (for example, the Wiener index), produces a more general screening approach than an invariant that considers this structural information (for example, the Overall Wiener index). Use of this screening level requires development of algorithms for construction of molecular graphs starting from the value of the utilized invariant. It is a



combinatorial problem whose complexity depends on the utilized invariant between other factors and which the authors have approached satisfactorily by use of heuristics that reduce the combinatorial explosion inherent to the problem.

On the basis of the described model in this article, we carried out construction of a screening tool, using of Java language, which allows their interoperability and easy distribution and whose description will be the task of the next article of this series.<sup>20</sup>

**Abbreviations:** CG cycle graph: graph in which the nodes represent the cycles and the edges represent the common nodes among those cycles present in a molecular graph *G*. CCG cycle and chain graph: graph in which the nodes represent the cycles and the chains and the edges represent the common nodes among those cycles and between the cycles and the chains present in a molecular graph *G*. CCG<sub>1</sub>: CCG in which the nodes that represent the chains of the graph *G* do not take into account the size of the chains. Chains of different size in the *G* graph are labeled with the same type of node in the CCG<sub>1</sub> graph. CCG<sub>X</sub>: CCG in which the nodes that represent the chains of the graph *G* take into account the size of the chains. Chains of different size in the *G* graph are labeled with a different type of node in the CCG<sub>X</sub> graph. ECG (equivalent cycle graph): graph in which the nodes represent the classes of equivalence of cycles and the edges represent the common nodes among those cycles present in a molecular graph *G*. Two cycles (or chains) of a *G* graph are equivalent, that is, they belong to the same equivalence class, if and only if the cycles are equal (equal number and type of nodes) and they participate in the same number and type of relationships with the remaining nodes of the *G* graph. ECCG<sub>1</sub> (equivalent CCG<sub>1</sub> graph): graph in which the nodes represent the classes of equivalence of the nodes (corresponding to cycles and chains) present in a CCG<sub>1</sub> graph. ECCG<sub>X</sub> (equivalent CCG<sub>X</sub> graph): graph in which the nodes represent the classes of equivalence of the nodes (corresponding to cycles and chains) present in a CCG<sub>1</sub> graph. TS: topological shape based on cycles and chains. A representation, at any level of abstraction, under the screening model proposed of the basic topological elements of the molecular structure. Each representation or TS represents a possible set of molecules which will be recovered from the database under that screening criterion. TS<sub>X</sub>: representation of the screening model in the cycle level. It is a TS in which only the cycles of the molecular structure are considered. TS<sub>C</sub>: representation of the screening model at the composition level. It is a TS in which the cycles and chains are considered (but not the size of the chains) of the molecular structure. TS<sub>S</sub>: representation of the screening model at the size level. It is a TS in which the cycles and the chains of the molecular structure are considered

## REFERENCES AND NOTES

- (1) Rouvray, D. H.; Balaban, A. T. Chemical Applications of Graph Theory. In *Applications of Graph Theory*; Wilson, R. J., Beineke, L. W., Eds.; Academic Press: New York, 1979; pp 177–221.
- (2) Cerruela García, G.; Luque Ruiz, I.; Gómez-Nieto, M. A. Representation of the Molecular Topology of Cyclical Structures by means of Cycle Graphs: 1. Extraction of Topological Properties. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (2), 447–461.
- (3) Luque Ruiz, I.; Cerruela García, G.; Gómez-Nieto, M. A. Representation of the Molecular Topology of Cyclical Structures by means of Cycle Graphs: 2. Application to Clustering of Chemical Databases. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1383–1393.
- (4) Kantardzic, M. *Data Mining: Concepts, Models, Methods, and Algorithms*; Wiley-IEEE Computer Society Press: New York, 2002.
- (5) Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 233–245.
- (6) Downs, G. M.; Barnard, J. M. Clustering and Their Uses. In *Computational Chemistry. In Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 2003; Vol. 18, pp 1–39.
- (7) Randic, M. Topological Indices. In *Encyclopedia of Computational Chemistry*; Schleyer, P., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; Wiley: Chichester, 1998; pp 3018–3032.
- (8) (a) Todeschini, R.; Consonni, V. The Handbook of Molecular Descriptors. In *The Series of Methods and Principles in Medicinal Chemistry*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-VCH: New York, 2000; Vol. 11, p 680. (b) Balaban, A. T.; Ivanciuc, O. Historical Development of Topological Indices. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999; pp 21–57.
- (9) Randic, M. Novel Shape Descriptors for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 607–613.
- (10) Cerruela García, G.; Luque Ruiz, I.; Gómez-Nieto, M. A. Cyclical Conjunction: An Efficient Operator for Extraction all Cycles in Graphs. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1415–1424.
- (11) Riley, D. *The Object of Data Abstraction and Structures*; Pearson Addison-Wesley: Reading, MA, 2002.
- (12) Xu, Y. J.; Johnson, M. A. Algorithm for Naming Molecular Equivalence Class Represented by Labelled Pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181.
- (13) Kier, L. B. Indexes of Molecular Shape from Chemical Graphs. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Scientific Publishing: New York, 1990; pp 151–174.
- (14) Randic, M.; Pompe, M. The Variable Molecular Descriptors Based on Distance Related Matrices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 575–581.
- (15) CODESSA (Comprehensive Descriptors for Structures and Statistical Analysis. <http://www.semichem.com>.
- (16) Li, X.; Lin, J. The Valence Overall Wiener Index for Unsaturated Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1358–1362.
- (17) Cerruela García, G.; Luque Ruiz, I.; Gómez-Nieto, M. A.; Cabrero Doncel, J. A.; Guevara Plaza, A. From Wiener Index to Molecules. Submitted to JCICS.
- (18) Figueras, J. Ring Perception Using Breadth-First Search. *J. Chem Inf. Comput. Sci.* **1996**, *36*, 986–991.
- (19) SPECS and BioSPECS B.V. <http://www.specs.net>.
- (20) Luque Ruiz, I.; Cerruela García, G.; Gómez-Nieto, M. A. Representation of the Molecular Topology of Cyclical Structures by means of Cycle Graphs: 4. A Screening Tool of Chemical Databases. Manuscript in preparation.

CI049889J