

Representation of the Molecular Topology of Cyclical Structures by Means of Cycle Graphs. 2. Application to Clustering 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, Building C2, Plant 3, E-14071 Córdoba, Spain

Received December 6, 2003

The great size of chemical databases and the high computational cost required in the atom–atom comparison of molecular structures for the calculation of the similarity between two chemical compounds necessitate the proposal of new clustering models with the aim of reducing the time of recovery of a set of molecules from a database that satisfies a range of similarities with regard to a given molecule pattern. In this paper we make use of the information corresponding to the cycles existing in the structure of molecules as an approach for the classification of chemical databases. The clustering method here proposed is based on the representation of the topological structure of molecules stored in chemical databases through its corresponding cycle graph. This method presents a more appropriate behavior for others described in the bibliography in which the information corresponding to the cyclicity of the molecules is also used.

1. INTRODUCTION

Nowadays, one of the main problems that the chemical scientific community (as well as in other areas) tries to solve is the handling of the immense quantity of information which they deal with. The size of the databases constantly grows until reaching millions of records, for which appropriate tools are required: (a) the organization and manipulation of the information and (b) the efficient access to this information.^{1–4}

For two decades, a great number of chemical databases have been developed which are the support of the chemical community in teaching and researching.^{5–11} The volume of these databases has grown to the point that it requires a tuning of the physical database structure in order to allow efficient access to the information stored.

To diminish the computational cost of this process data mining techniques^{12–14} are applied implying, among other actions, the proposal of models that allow the classification of the database elements in clusters or classes and the development of searching and recovering algorithms of the information.

But, evidently, the adaptation of the clustering method is dependent on the finite set of approaches that will be used in the later search process. So, if the database is classified in the function of a property P_1 (i.e. molecular weight, boiling point, Wiener index, etc.) a search process in which a property $P_2 \neq P_1$ is considered as a search approach will have a poor performance.

Another patent problem is the size of the set of recovered records. If the cardinality of this set is high, the later process of comparing the search pattern atom to atom with each of the recovered records supposes a high cost. However, if the cardinality of the set is very low, it is possible that in the screening process database elements of interest have been left without being recovered.

From an abstract point of view a clustering model supposes a generalization model.^{14–16} Individual elements (molecules) are generalized in an object class (a cluster) using an ownership function (one or a set of approaches or classification properties). Therefore, the database clustering supposes “a certain loss” of information, since all generalization supposes loss of specialized information. Although the clustering introduces the advantage of diminishing the initial size of the information domain on which a search is carried out, it is done with the representatives of the clusters instead of each element of the database.

For everything that is exposed, it becomes necessary to propose new clustering models guided to the development of physical structures of the chemical databases that allow the recovery of information that take into account, at least, the set of criteria more used or of more interest for the chemists.

The information corresponding to the present cycles in the topological structure of the molecules is one of these approaches. It is well-known that the presence and characteristic of the cycles in a molecule determine their physical, chemical properties, and biological activity. In fact, cycles exist in the molecular structure of most chemical compounds of pharmacological interest.

The cyclicity of the molecules has been used previously for the proposal of clustering models,^{17,18} although in these approaches partial information only is used and slanted regarding the molecular cyclicity. In a previous paper¹⁹ we have proposed that, since the computational cost of the obtaining of all the cycles of a molecule is acceptable, the use of this information in computational chemistry is convenient.

Thus, knowing all the cycles in the topological structure of a molecule it is possible to build an isomorphic representation of the molecular structure denominated *Cycle Graph*. The cycle graph has been described in a previous paper,¹⁹ and it consists of a weighted, colored, and nondi-

* Corresponding author phone: +34-957-212082; fax: +34-957-218630; e-mail: ma1lurui@uco.es.

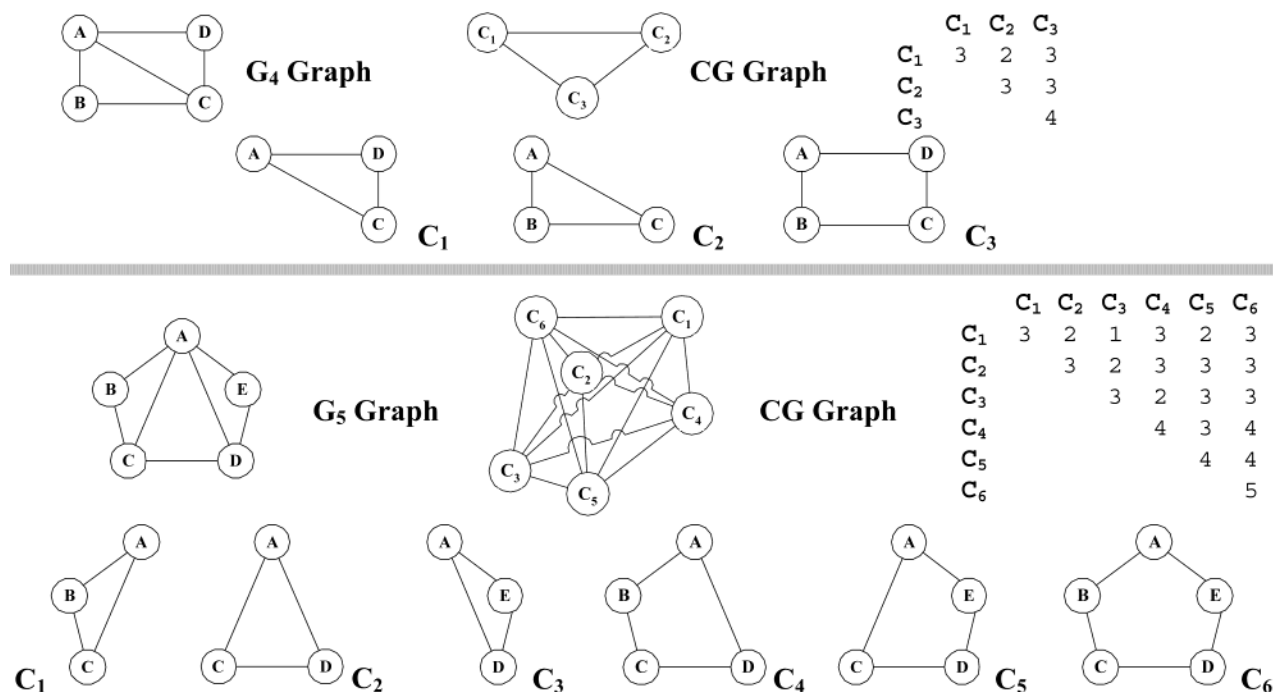


Figure 1. Example of CG graphs for two cyclical structures.

rected graph in which the nodes represent each one of the cycles and the edges represent the common nodes among the cycles in the molecular structure.

In this paper we present a new clustering method based on the cycle graph for the classification of chemical databases. The clustering method is evaluated for different databases and compared with the methods described in the bibliography that also make use of the information of the molecular cyclicity demonstrating the usefulness of the proposed method.

Section 2 gives a brief description of the cycle graph and of the equivalent cycle graph which allow us to represent cyclic molecular structures, and in section 3 this representation is enlarged to structures in which the acyclic chains are considered. In section 4 the clustering proposed method is described, and in section 5 the results obtained are presented. Last, these results are discussed comparing them with those described in the bibliography.

2. REPRESENTATION OF CYCLICAL STRUCTURES THROUGH CYCLE GRAPHS

The topological structure of molecules is usually represented by means of a molecular graph G whose nodes represent the atoms and the edges represent the bonds in the molecule. Due mainly to the size of these graphs—the size of the molecules—has intended a great number of isomorphic representations of the molecular graph.^{17,20,21}

A homomorphic graph G' of a molecular graph G is a graph, generally smaller than G , that is suitable for the study of certain properties corresponding to the molecule that G represents, and it can be used conveniently to characterize G , but in the reduction process undoubtedly some properties are lost of the molecule represented by the graph G which has taken its place.^{16,19}

The authors have proposed the construction of an isomorphic graph G' based on the information corresponding to all

the present cycles in the G graph, denominated Cycle Graph (CG).

Given a cyclical structure represented by a molecular graph G , all the present cycles in the graph can be extracted efficiently using the algorithms proposed by the authors.^{22,23} Knowing the set of cycles C present in a molecular graph G , a homomorphic graph of cycles CG can be built with the following characteristics:

□ The CG graph is a nondirected, weighted, and colored graph.

□ The CG graph has the same number of nodes as elements which 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.

□ The edges of the CG graph are labeled, and they represent the relationships among cycles of the C set. Given two nodes $i, j \in CG$, that 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 CG graph can be represented by means of a symmetrical weight matrix W_{CG} whose elements $W(i, i)$ take a value equal to the number of nodes of G that participate in the C_i cycle, and any element $W(i, j)$ is equal to the number of common nodes in the G graph among the cycles C_i and C_j .

Figure 1 shows an example of the CG graph and their corresponding weight matrix for two cyclical structures. As can be observed, the nodes number of the CG graph can be higher (G_5), lower (G_4), or equal to the number of nodes of the G graph. For complex cyclical structures, with a high number of interrelated cycles (K_n graphs), the CG graph is more complex than the G graph, due to the existence of a high number of complex cycles composed of some elemental cycles that are used properly for the extraction of topological

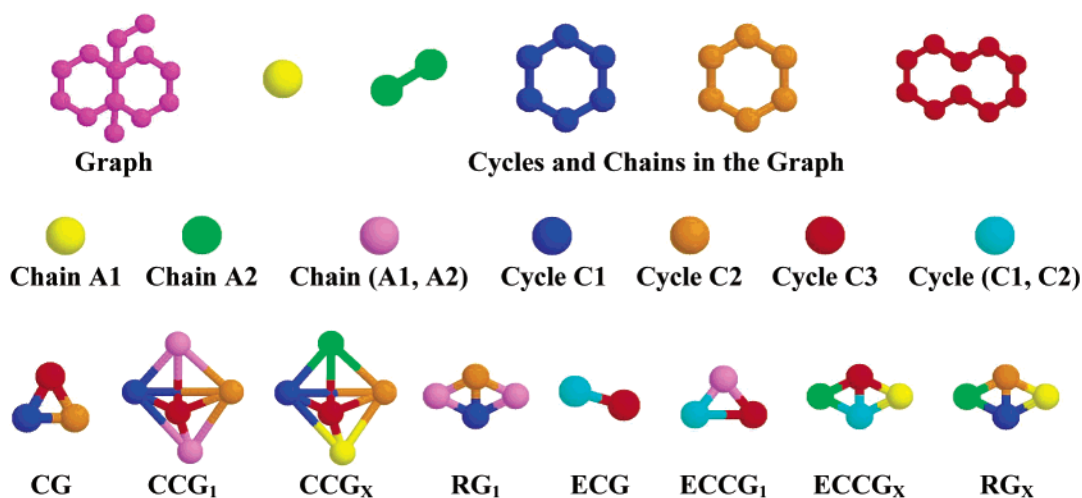


Figure 2. Molecular structures and their corresponding graphs for all models described in the paper.

properties and the clustering of chemical databases as we will describe later in the manuscript.

2.1. Equivalent Cycle Graphs. The concept and use of class of equivalence in graph theory and, mainly, in its application to the molecular graphs is very extended, its efficiency being demonstrated in many cases.^{24,25} In the *CG* graph it is considered that a class of equivalence is formed by the following:

1. Nodes of *CG* graph with the same color and therefore representing cycles composed of the same number of atoms.
2. Nodes of the *CG* graph participating in the same number and type of relationships.

On occasion it can be convenient to carry out a new step in the generalization process for which the cycle graph has been built. In this step the equivalent cycle graph is built *ECG*,^{19,24} a homomorphic graph to the *CG* graph, in the following way:

- The *ECG* graph is a nondirected, weighted, and colored graph.
- The *ECG* graph is composed of as many nodes as equivalent classes of cycles present in *CG* graph.
- The edges of the *ECG* graph are labeled and represent the relationships—number of common nodes—among the equivalent class of cycles in *CG*.

3. CYCLE AND CHAIN GRAPHS

In molecules where all the atoms form part of a cycle (i.e. benzenoid systems) the use of the graphs *CG* and *ECG* do not cause a loss of information (besides its own due to the generalization process), but in molecules in which atoms that are not part of cycles exist these representations introduce a high reduction.

This reduction can be conveniently used in clustering processes in which it is interesting that molecules such as C_xH_x , $m-C_xH_x$, $e-C_xH_x$, $p-C_xH_x$, $m,p-C_xH_x$, in general: $(a_i, b_j, \dots, z_k)-C_xH_x$, where a , b , z , representing aliphatic chains, are assigned to the same cluster (class), to only classify the database based on the cyclical structures present in the molecule. However, when it is of interest to classify based on the general topology of the molecules (cycles and chains), it becomes necessary to enlarge the proposed model.

The cycle and chain graphs *CCG* and the equivalent cycle and chain graphs *ECCG* are built in a similar way to the *RG*

graphs (*Cyclic Information Reduced Graph*).¹⁷ 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 which 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 the *CCG* graph, obtaining the classes of equivalence existing in *CCG* (taking into account that the nodes corresponding to cycles are colored in different ways to the nodes corresponding to chains) and carrying out the same process aforementioned for the construction of the *ECG* graph.

Figure 2 shows a molecule example and its corresponding graphs *CG*, *ECG*, *CCG*, and *ECCG*, and we can observe the characteristics of the described representation models. In Figure 2 the representation of the *RG* graph corresponding to the same molecules are also shown, highlighting the differences between this representation and the proposals in our work.

The *RG* graph does not color the nodes corresponding to the chains, which implies a reduction (hiding) of information in the isomorphism process. Also, the number of nodes of the *RG* graph is in most cases minor to that in the *CCG* graph, because the *RG* graph only considers the cycles of the *SSCE* set (*Set of Smallest Cycles at Edges*)—equal or higher by an unit to the *SSSR* set—, while in the *CCG* graph (as in the *CG* graph) all the cycles in the *G* graph are considered.

4. CLUSTERING OF CHEMICAL DATABASES BASED ON CYCLE GRAPHS

Under the cycle graphs model for the representation of molecular structures of the chemical compounds, the clustering process of the chemical databases is calculated using the following steps.

4.1. Preprocessing. The preprocessing stage consists of generating the representation based on the cyclicity of the molecular structure of the database elements. For the study and evaluation of the clustering method proposed we have generated for each element of the database the following:

- The corresponding *CG* graph.

Table 1. Characteristic of the Different Clustering Models for a Database of 20 677 Compounds^a

model	I _N	T _N	C _T	N/CN _M	#N/CNm	PC _M	NPC _M	PC _m	#CPC _m	PN _M	NPN _M	PN _m	#NPN _m
CG	1–13	13	498	6/70	1/5	2289	3	1	183	4673	3	7	1
ECG	1–13	13	371	4/73	1/1	5472	2	1	144	5884	2	1	1
CCG _I	2–19	16	4185	10/676	2/1	497	5	1	2121	3420	7	1	2
ECCG _I	2–14	13	2337	9/379	1/7	1137	5	1	1073	3564	7	25	1
CCG _X	2–19	16	14092	7/2344	2/1	38	6	1	10792	3420	7	1	2
ECCG _X	2–14	13	11145	7/2047	1/38	61	7	1	7659	3581	7	54	1
RG _I	2–14	13	4185	8/829	1/4	497	5	1	2121	4206	7	12	1
RG _X	2–14	13	14092	7/3015	1/11	38	6	1	10793	4206	7	12	1

^a I_N: interval of values of the number of nodes. T_N: number of nodes with population. C_T: number of total classes. N/CN_M: node with the maximum number of classes/number of classes. #N/CNm: number of nodes with the minimum number of classes/number of classes. PC_M: population of the class with maximum population. NPC_M: value of the node corresponding to the PC_M. PC_m: population of the class with minimum population. #CPC_m: number of classes corresponding to the PC_m. PN_M: population of the node with maximum population. NPN_M: value of the node corresponding to the PN_M. PN_m: population of the node with minimum population. #NPN_m: number of nodes corresponding to the PN_m.

- The corresponding CCG graph for each one in the following representations:

- Labeling the nodes corresponding to the chains with the size of the same ones (number of atoms). This graph is named CCG_X.

- Labeling the nodes corresponding to the chains with the same value independent of the size of the chain. This graph is named CCG_I, so this representation does not take into account the different size of the chains existing in the G graph.

- The corresponding ECG graph.
- The ECCG graph, in each one of the previously described representations (ECCG_X and ECCG_I).

Also to validate the results of the proposed method we have carried out for each element of the database the following:

- The construction of the RG graph, in each of the following ways:

- Labeling the nodes corresponding to the chains with the size of the same ones, this graph is named RG_X.

- Labeling the nodes corresponding to the chains with same value independent of the size of the chain, as described in ref 16, this graph is named RG_I.

4.2. Processing. Once the corresponding representations are built the generated graphs are compared in order to find the isomorphic graph. In this process the Ullman algorithm²⁶ is used.

The graphs that are equal in the comparison process are assigned to one class or cluster. Two graphs are equal if they have the same number of nodes, of the same type (of cycles and chains), with the same label (size, for those representations in which this information is considered), and maintain the same set of relationships among them.

As result of this process an index is built with the following information:

- Identification of the molecule or database element.
- Nodes number of the graph in the study (each one of the representations described previously), with specification of the total nodes, nodes corresponding to cycles, and nodes corresponding to chains.

- The class (key) to which the molecule has been assigned.

This index allows a clustering of the database to be carried out in function of the following parameters:

- The number of classes or clusters generated. That is, the number of different graphs.

- The nodes number of the corresponding cycle graph. This level includes the previous one since the structures

assigned to the same class should have equal number, type, labels, and relationships among the nodes, although structures with an equal number of nodes can be assigned to different classes.

5. ANALYSIS OF THE RESULTS

The evaluation of the proposed clustering methods has been carried out on several databases of public domain of different size and characteristics. In the clustering process, the different representations described in the previous section have been generated, building the corresponding index and carrying out the clustering of the database.

The evaluation process has been carried out considering the following factors:

- Number of classes or clusters generated for each clustering model.

- The population in each cluster (elements of the database assigned to the cluster).

- The population's distribution in each cluster. The calculation of this parameter is made using the algorithm for the calculation of the similarity proposed by the authors,²⁷ the similarity corresponding to each couple of elements of the database assigned to the cluster is calculated, and the similarity average is also calculated.

For each clustering model we have extracted a set of statistic descriptors (see Table 1) that will allow us to analyze properly the usefulness of the proposed models and that is described later, showing the results on a database of 20677 records. The database has been extracted from SB2C_20T containing synthetic organic compounds with a wide molecular diversity.⁸

Figure 3 shows the distribution of the elements of the database for the different clustering models in function of the nodes number and the generated classes. The classes have been numbered consecutively in function of the nodes number of the corresponding graph (model). A cluster is identified by the couple: nodes number of the graph, class number (set of equal graph—using the Ullman similarity algorithm—and different to another set of graphs with equal or different number of nodes).

5.1. Consideration of the Classes of Equivalence. The interval of nodes number I_N determines the number of classes that will be generated directly. As the number of nodes increases the number of classes also increases, diminishing the class population. The representations based on the

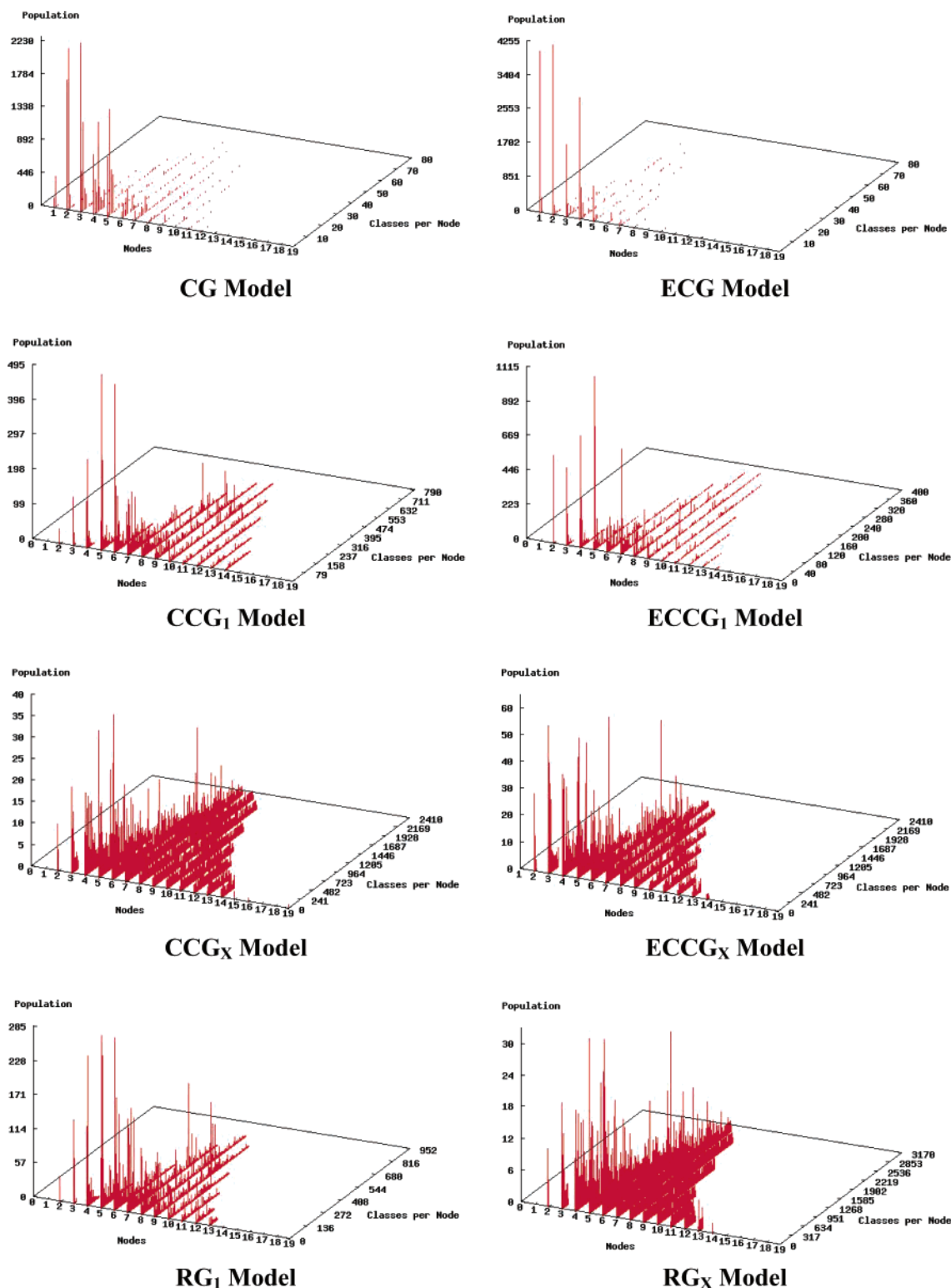


Figure 3. Behavior of the different clustering models regard to the number of nodes and clusters by node.

equivalent cycles diminish the nodes number since different graphs *CG* (or *CCG*) can give place to equal graph *ECG* (or *ECCG*).

As observed in Table 1 and Figure 3, there do not exist populations (molecules) for all I_N values. The T_N parameter informs the nodes number (in the interval I_N) for which some element of the database in the different representation models exists. We can see that T_N has the same behavior commented previously for I_N .

The number of total classes also depends on the representation used for the clustering process. As observed in Table 1 and Figure 3, the number of clusters C_T diminishes with the use of the graphs based on equivalent cycles, diminishing the maximum number of classes for a given node (CN_M). In the *CCG₁* model the grouping is carried out for graphs of 10 nodes, although with a smaller discrimination (676 classes). This is because *CCG₁* graphs with 9, 10, and more nodes are reduced to *ECCG₁* graphs with 9 nodes due

to its composition of equivalent cycles and chains. Notice that in the CCG_I and $ECCG_I$ models the chains size is not considered.

In some clustering models structures exist that for a given number of nodes only exists in one class (CN_m) and with a population equal to 1 (PN_m). It is observed that this number of nodes ($\#N$ and $\#NPN_m$) diminishes again with the use of the representations based on the equivalent cycles (and chains), again due to the topological reduction that these representations imply faced with those based on the cycles (and chains).

This topological reduction means that in the models based on the equivalent cycles (and chains) the population of the classes increases, an effect that is observed in the parameter PC_M that informs of the class population with higher population, which happens for some values of nodes (NPC_M), and diminishes at the same time the number of classes with a minimum population ($PC_m = 1$), as can be seen in the values of $\#CPC_m$.

The analysis of these results shows that the models based on the equivalent cycles (and chains) give place to a better grouping than those based on the cycles (and chains) since they cause a decrease in the number of nodes and classes with minimum populations and a decrease in the number of total classes with an increase of population of each class.

5.2. Consideration of the Presence of Chains. The consideration of the acyclic chains in the clustering models gives in all cases an increase in discrimination, and, therefore, an increase in the number of classes. This effect means a decrease in the class population due to the discrimination produced by the consideration of the chains in the structure of the chemical compounds.

Although the number of classes is practically multiplied by 8 (see the values of C_T for the CG and CCG_I models in Table 1) when considering the chains, the population of these classes diminishes in an order of 75%, an effect due to a better distribution of the populations and to the increase of the nodes and classes with minimum population.

The models based on acyclic chains produce a displacement of the graph grouping to nodes with a high value of the node number and an increase in the distribution of the graphs in classes, also increasing the number of classes with minimum population.

5.3. Consideration of the Chains Size. The models that consider the size of the chains (CCG_X and $ECCG_X$), compared with those that do not consider it (CCG_I and $ECCG_I$), produce an increase in the discrimination as is appreciated in the parameter C_T in Table 1 and Figure 3. This increase in the number of total classes is more marked in the CCG_X model than in the $ECCG_X$, since most of the chemical compounds have cycles of close size (rings of 5 and 6 atoms mainly), while the main difference is in the size of the chains.

In both cases (CCG_X and $ECCG_X$) the population of the classes diminishes, decreasing more patiently in the CCG_X model. So, the number of classes with a minimum population (equal to 1) of the $ECCG_X$ model is lower than the CCG_X model. This effect is because the CCG_X model improves the discrimination among the topological structures when considering the size of cycles and chains. This discrimination is reduced in the $ECCG_X$ model due to the topological

reduction that takes place when considering the equivalent cycles and chains.

5.4. Comparison with the RG Model. The RG_I model¹⁷ considers the set of cycles $SSCE$ and the present chains in the molecular structure, not considering the chain size.

As shown in Table 1, under this model the number of nodes is inferior to some of the models proposed in this work ($I_N = 14$ and $T_N = 13$) which should be translated, in principle, to a smaller number of classes. However the number of total classes ($C_T = 4185$, $CN_M = 829$) is higher in some cases for our models and equal to the CCG_I model.

Although for the RG_I model the total number of nodes is lesser than in some of our proposed models, the RG_I model presents a number of classes with a minimum population (equal to 1) close to and even superior to some of our proposals, which gives a place to a grouping of the elements of the database in certain nodes (e.g. $NPN_M = 7$).

The behavior of the RG_I model is very close to the CCG_I model (see Table 1); however, as the CCG_I considers all the cycles, it distributes the molecules in a higher interval of nodes number than the RG_I model, although the classes are composed of the same molecules.

5.5. Modification to the RG_I Model. The RG_X Model. With the purpose of improving our study of the clustering models based on the information of the cyclicity in molecular graphs, we have carried out a modification to the model proposed by Dury.¹⁷ The RG_X model is close to the RG_I , but in this case the size of the chains coloring the corresponding nodes in the graph is considered.

Evidently, as is observed in Table 1, the RG_X model does not introduce any modification of the parameters I_N and T_N , since no operation on the corresponding RG graph occurs. However, the RG_X model gives place to some improvements with regard to the RG_I : an increase in the number of classes (C_T), the populations of the classes decrease (PC_M), higher discrimination of the "similar" elements of the database (CN_M). But also there are some inconveniences such as the increase in the number of classes with minimum population ($\#CPC_m$), which is due to the higher discrimination that this model introduces.

The RG_X model has the same behavior with regard to the CCG_X as previously described for the RG_I model with regard to the CCG_I model. The CCG_X model distributes the class of the RG_X model in a higher number of nodes (see Table 1), although the classes are composed of the same molecules.

5.6. Characteristics of Grouping of the Different Models. Each one of the grouping models studied in this work makes use of the molecular topology and especially of the cyclicity to carry out the classification of the chemical database in classes or clusters.

To evaluate these models an analysis of the topological similarity of the molecules that are assigned to each cluster has been carried out. This end has been obtained by means of the cosine similarity index²⁹ and making use of an algorithm developed by the authors.²⁷ So, the topological similarity among all the molecules assigned to each class is calculated, being observed as the following:

- The ECG model generates clusters in which the molecules have a lower structural similarity than the CG model. This fact is due to the topological reduction that occurs when extracting the classes of equivalent cycles from the cycle graph, which causes a reduction in the number of

Table 2. Statistic Parameters Obtained for the Different Clustering Models on a Database of 20 677 Records^a

model	clusters/population					nodes/population					nodes/classes				
	%S _C	%D _C	AP _C	CE _C	ENC _C	%S _N	%D _N	AP _N	CE _N	ENC _N	%S _S	%D _S	AP _S	CE _S	ENC _S
CG	36.7	13.3	41.5	5.4	43	0.0	0.0	1590.5	2.8	7	0.0	0.0	38.3	0.2	1
ECG	38.8	12.9	55.7	4.0	16	7.7	0.0	1590.5	2.5	6	7.7	7.7	28.5	0.2	1
CCG _I	50.7	16.5	4.9	10.3	1253	12.5	0.0	1292.3	3.3	10	12.5	0.0	261.6	1.1	2
ECCG _I	45.9	17.0	8.9	8.6	377	0.0	0.0	1590.5	3.2	9	0.0	0.0	179.8	0.7	2
CCG _X	76.6	14.2	1.5	13.5	11287	12.5	0.0	1292.3	3.3	10	12.5	0.0	880.8	2.6	6
ECCG _X	68.7	16.2	1.9	12.9	7472	0.0	0.0	1590.5	3.2	9	0.0	0.0	857.3	2.2	5
RG _I	50.7	16.5	4.9	10.3	1253	0.0	0.0	1590.5	3.1	8	0.0	0.0	321.9	1.1	2
RG _X	76.6	14.2	1.5	13.5	11287	0.0	0.0	1590.5	3.1	8	0.0	0.0	1084.0	2.4	5

^a %Single: percentage of singletons, %Double: percentage of doubletons, AP: average population of the clusters, CE: clustering entropy, ENC: effective number of clusters.

classes (see Table 1) and, because none of these models considers the chains, there being a better distribution of the molecules in the CG model than in the ECG.

- The models based on the consideration of cycles and chains (CCG) generate clusters with a higher similarity among the molecules than the model based only on cycles. The reason is obvious; the consideration of the chains gives place to a higher specialization, which is translated to an increase in the clusters number (see Table 1) in which the assigned molecules are more similar structurally.

- The models based on the equivalent cycles and chains (ECCG) produce clusters where the assigned molecules are less similar than in the models based on the cycles and chains (CCG). Since the ECCG model generates fewer clusters than the CCG models that is, the grouping model is more general, the clusters are formed by molecules that are more structurally diverse, which is shown in the smallest clusters number generated in the ECCG models with regard to the CCG.

- The consideration of the chains size in the ECCG_X model produces a very appreciable effect with regard to the ECCG_I model. As it is observed in Table 1 an increase in the number of classes is observed and, therefore, in the similarity of the molecules assigned to each class.

- The CCG_X (or RG_X) model gives place to clusters where the molecules are much more similar than in the CCG_I (or RG_I) model. The consideration of the chains size generates a higher clusters number and, therefore, a higher specialization in the classification process.

Figure 4 shows by means of histograms the clusters populations in function of the nodes number of the corresponding graphs for the different studied models. Also shown in Figure 4 is the discrimination power (in the interval 0–100) for each value of the nodes number, obtained as the ratio between the number of clusters and the population by node.

As can be appreciated in Figure 4, the power of discrimination of the ECG model is very close to the CG despite the decrease of the clusters number that the ECG model produces.

However the discriminatory power of the ECCG_I model is sensibly inferior to the CCG_I. As is appreciated in Table 1, the ECCG_I model produces a decrease of around 50% of the clusters number that generates the CCG_I model. This fact gives place to a higher diversity of the clusters populations, due evidently to the consideration of the equivalent cycles and chains (without take into account the size) which produces similar ECCG_I graphs among molecules with enough diversity.

However the discriminatory power of the ECCG_X model is not very sensibly inferior to the CCG_X. As is appreciated in Table 1, the ECCG_X model produces a poor decreasing of the clusters number that generates the CCG_X model. This fact is due to the consideration of the chain size which gives places to a low grouping when equivalent cycles and chains are considered, and, therefore, the clusters of the CCG_X and ECCG_X models are composed of a close number and type of molecules.

When the CCG_X (or RG_X) models are analyzed, the highest discriminatory power is observed. It is evident that this model produces a higher number of clusters than the remaining models, and, therefore, these clusters are composed of few and very similar molecules although the CCG_X model distributes the clusters better than the RG_X model over a wide number of nodes. The same behavior is observed for the CCG_I and RG_I models.

6. DISCUSSION

The usefulness of the clustering methods is usually measured by the calculation of the entropy that produces the classification process.²⁹ The entropy of the clustering is calculated by means of the expression:

$$CE = - \sum_{i=1}^q f_i \log_2 f_i \quad (1)$$

where q is the clusters number and $f_i = n_i/N$ is the population frequency in each cluster, calculated as the ratio among the population of each cluster (n_i) and the number of elements of the database (N).

Knowing CE , the effective number of clusters can be calculated with the expression:

$$ENC = 2^{CE} \quad (2)$$

Table 2 shows the values of these parameters (Clusters-Population) as well as the singletons percentage (%S_C)—clusters number with population equal to 1—the doubletons percentage (%D_C)—clusters number with population equal to 2—and the average of the population (AP_C)—ratio between the elements of the database and the clusters number.

Also shown in Table 2 are these same parameters obtained as follows:

- Nodes-population: where q is equal to the nodes number, n_i is equal to the total population assigned to each value of q , AP_N is equal to the ratio among the elements of the

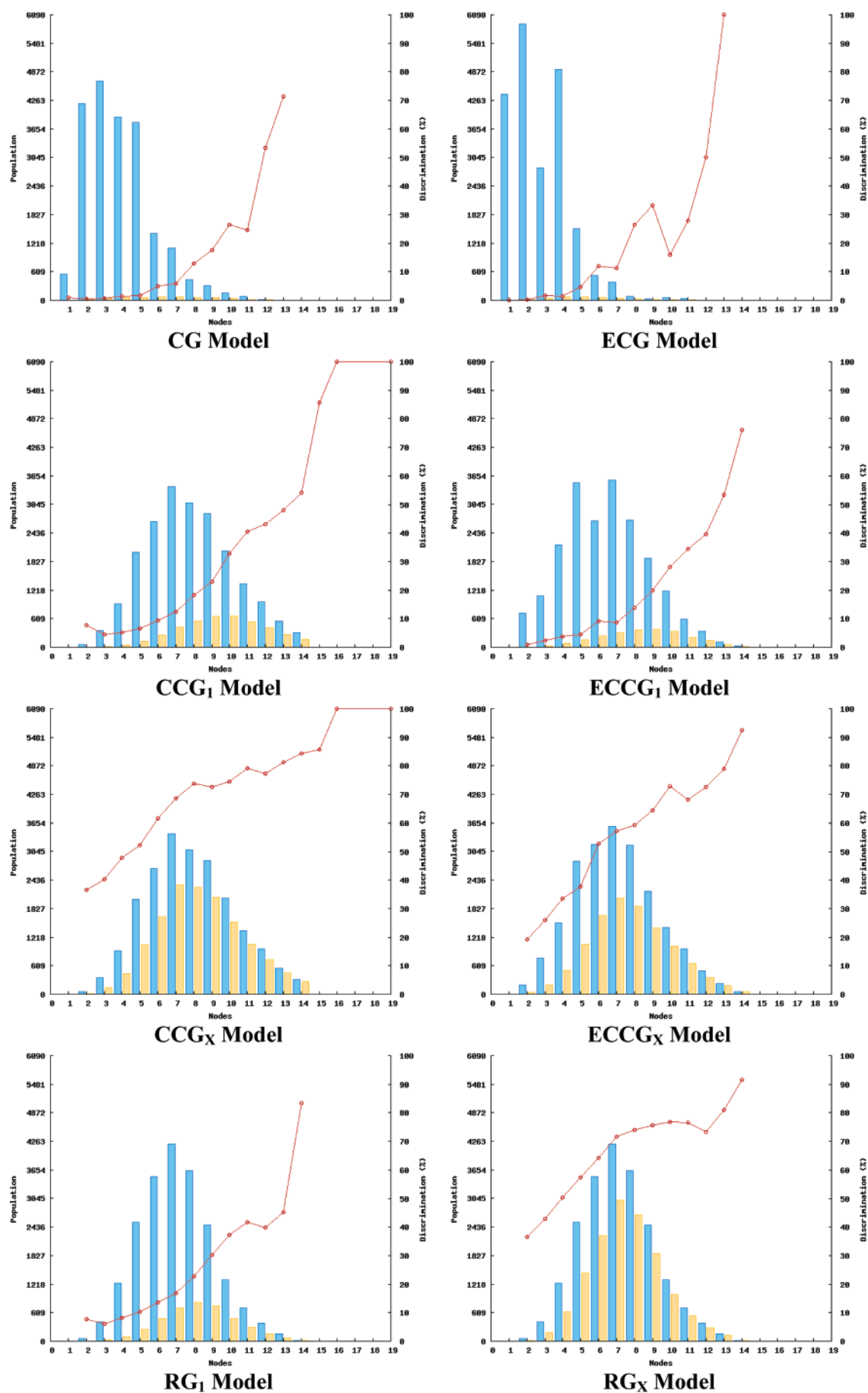


Figure 4. Histograms showing the behavior of the different clustering models regard to the number of nodes: blue bars, population by node; yellow bars, number of clusters by node; red lines, discrimination power.

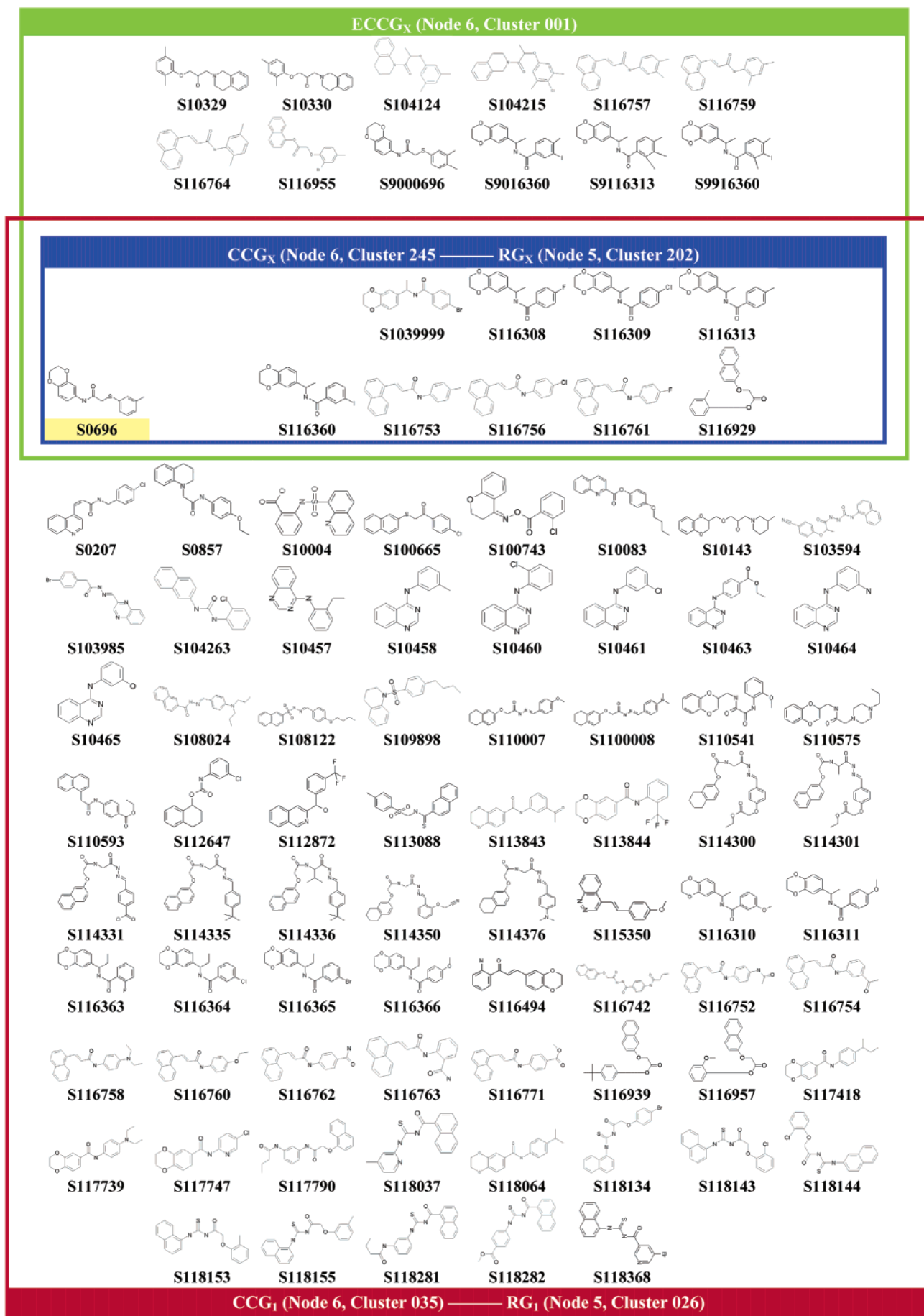


Figure 5. Classification of a S0696 molecule making use of the ECCG_X, CCG_X, RG₁, and RG_X models.

database and the nodes number, and $\%S_N$ and $\%D_N$ are the percentages of the nodes number with a population equal to 1 and 2, respectively.

• *Nodes-classes*: where q is equal to the nodes number, n_i is equal to the classes number for each value of q , N is the clusters number generated in the classification process, AP_S is equal to the ratio among the clusters number and the number of nodes, and $\%S_S$ and $\%D_S$ are the percentages of the classes number with a population equal to 1 and 2, respectively.

The information shown in Table 2 corroborates the analysis of the results described in the previous section. It can be seen that the highest values of CE are presented for the models CCG_X , $ECCG_X$, and RG_X , for which a higher value of ENC is obtained.

It is observed that for all models the singletons and doubletons are due to nodes in which one and two elements have been assigned respectively to one or two classes (equal values for $\%S_N$, $\%D_N$ and $\%S_S$, $\%D_S$). In all cases this behavior is due to graphs (molecules) in which a high number of cycles exist (rarely due to the number of chains), because the number of molecules with these characteristics is low in not very large chemical databases.

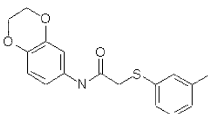
The values of $\%D_C$ are close for all models; however, $\%S_C$ increases with the specialization of the models. So, the consideration of acyclic chains (GCC versus GC) increase the singletons, increasing the singletons more when the chain size is taken into account (GCC_X vs GCC_I or RG_X vs RG_I). The models based on equivalent cycles and chains give a smaller number of singletons due to the class grouping produced.

The values of $\%S_N$ and $\%D_N$ (or $\%S_S$ and $\%D_S$) are zero for those models considering a smallest number of nodes of the corresponding generated graphs. Although population averages by nodes (AP_N) are very close for all models, the class averages by nodes (AP_S) increase when cycles and chains are considered (CCG_I vs CG), also with the chain size (CCG_X vs CCG_I), diminishing for the models based on equivalence classes ($ECCG_I$ vs CCG_I or $ECCG_X$ vs CCG_X). This effect is due to the models considering that acyclic chains distribute the graphs in a higher number of nodes, producing very different graphs which are assigned to different classes. This behavior is corroborated by the population average of the clusters (AP_C). This value diminishes when cycles and chains are considered and also with the consideration of chains size.

The maximum discrimination powers, as expected, are obtained for the models CCG_X and RG_X , for which are generated a higher clusters number (see Table 1) as well as higher values of CE_C and ENC_C , with a percentage of effective clusters of the order of 80%, although the CCG_X model presents a higher value of ENC_N and ENC_S and a lower value of AP_N and AP_S contributing a better classification of the database elements.

The distribution of the database elements in the clusters for the different models is appreciated with the example shown in Figure 5 and Table 3. In Table 3 the classification of a molecule of the database to the corresponding cluster is shown, besides the population of the cluster, the average of the similarity among the molecules assigned to the clusters, for the different clustering models treated in this paper.

Table 3. Example of Distribution in Clusters for All the Proposed Models

compound	model	node/class	population	S_A
	CG	4/001	815	0.5944
	ECG	3/001	1815	0.5552
	CCG _I	6/035	79	0.6989
	ECCG _I	6/001	160	0.6523
	CCG _X	6/245	10	0.7862
	ECCG _X	6/001	22	0.7485
	RG _I	5/026	79	0.6989
	RG _X	6/202	10	0.7862
S0696				

We can observe in Table 3 the smallest populations in the clusters for the representations based on cycles and chains with regard to those only based on cycles, which translates to an increase of the average of the similarity of the cluster. Likewise the decrease of the cluster populations is observed in the representations based on the cycles (and chains) with regard to equivalent cycles (and chains), meaning an increase in the average of the similarity of the cluster.

We can see again the higher discrimination of the models CCG_X and RG_X , which generates clusters with low populations and with a very high similarity among the assigned compounds, while other models, such as CCG_I and $ECCG_X$, present the average of population clusters with an acceptable similarities average.

Figure 5 shows, for a molecule of the database, the elements of the cluster assigned for the $ECCG_X$, CCG_X (and RG_X), and CCG_I (and RG_I) models; in this case as in all cases both models (CCG_I – RG_I and CCG_X – RG_X) generate equal clusters (although assigned to different nodes). Figure 5 shows the hierarchical characteristics in the clustering process of the proposed models. So, the $ECCG_X$ model generates clusters including the clusters generated by the CCG_X model as well as other equivalent graphs. The CCG_I (RG_I) model has similar behavior with respect to the CCG_X (RG_X) model.

Throughout the paper we have presented a series of representation models based on the cyclicity of the structure of the chemical compounds as an approach for the construction of clustering models of chemical databases. These clustering models present different behavior in function of the generalization considered in the representation model. So, the models that consider the presence of the acyclic chains and their size offer a higher discrimination than the models that only consider the presence of cycles (and/or not) the presence of chains without considering the size of these.

Evidently, a very selective classification is not always favorable, since many clusters are generated with low populations, although in these cases the similarity of the molecules assigned to each cluster is very high. But neither is it convenient that the clustering model generates few classes with a high population, as the later comparison (atom to atom) of a molecule problem with the elements of the cluster will be very expensive.

Since the obtaining of the graph corresponding to each representation model for each element of the database is not very expensive once the CG graph is obtained, the clustering models tried in this paper present the advantage that it is feasible for the classification of a chemical database to use one or several of the proposed models, obtaining a multiple index with information of the cluster (node and class) to which each compound is assigned for each model. This

multiple index would allow the development of screening procedures with a high performance whose proposal will be the object of a future paper.

REFERENCES AND NOTES

- (1) Gillet, V. J.; Downs, G. M.; Ling, A.; Lynch, M. F.; Venkataran, P.; Wood, J. V.; Dethlefsen, W. Computer Storage and retrieval of Generic Chemical structures in Patents. 8. Reduced Chemical Graph and Their Applications in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, 27, 126–137.
- (2) Butina, D. Unsupervised data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated way to Cluster Small and Large data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, 39(4), 747–750.
- (3) Turner, D. B.; Tyrrel, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 18–22.
- (4) 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.
- (5) Daylight Chemical Information Systems Inc. <http://www.daylight.com>.
- (6) Molecular Simulations Inc. (Accelrys). <http://www.accelrys.com>.
- (7) MDL Information Systems Inc. <http://www.mdl.com>.
- (8) SPECS and BioSPECS B.V. <http://www.specs.net>.
- (9) NCI Drugs Information system. <http://cancer.gov/cancerinformation>.
- (10) MedChem/BioByte. <http://iris.pomona.edu/>
- (11) Chemical Abstract Service. <http://www.cas.org/>
- (12) Du, Y.; Liang, Y.; Yun, D. Data Mining for Seeking an Accurate Quantitative Relationship between Molecular Structure and GC Retention of Alkenes by Projection Pursuit. *J. Chem. Inf. Comput. Sci.* **2002**, 42(6), 1283–1292.
- (13) Cundari, T. R.; Russo, M. Database Mining Using Soft Computing Techniques. An Integrated Neural Network-Fuzzy Logic Genetic Algorithm Approach. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 281–287.
- (14) Kantardzic, M. *Data Mining: Concepts, Models, Methods, and Algorithms*; Wiley-IEEE Computer Society Pr, 2002.
- (15) Kaufman, L.; Rousseeuw P. J. *Finds Group in Data: An Introduction to Clustering Analysis*; John Wiley & Sons: 1990.
- (16) Riley, D. *The Object of Data Abstraction and Structures*; Pearson Addison-Wesley: 2002.
- (17) Dury, L.; Latour, T.; Leherter, L.; Barberis, F.; Vercauteren, D. P. A new graph Descriptor for Molecules Containing Cycles. Application as Screening Criterion for Searching Molecular Structures within Large Databases of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1437–1445.
- (18) Lipkus, A. H. Exploring Chemical Rings in a Simple Topological-Descriptor Space. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 430–438.
- (19) 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.
- (20) Thierauf, T. *The Computational Complexity of Equivalence and Isomorphism Problems*. Lecture Notes in Computer Science, 1852. Springer-Verlag: 2000.
- (21) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 338–345.
- (22) 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.
- (23) Cerruela García, G.; Luque Ruiz, I.; Gómez-Nieto, M. A. Parallel Algorithms for Graph Cycle Extraction Using the Cyclical Conjunction Operator. *J. Chem. Inf. Comput. Sci.* **2002**, 42(6) 1398–1406.
- (24) Alpin, J.; Mubarakzianow, R. The bases of Weighted Graphs. *Discrete Mathematics* **1997**, 1–11.
- (25) Rucker, G.; Rucker, C. Computer Perception of Constitutional (Topological) Symmetry: TOPSYM, a Fast Algorithm for Partitioning Atoms and Pairwise Relations among Atoms into Equivalence Classes. *J. Chem. Inf. Comput. Sci.* **1990**, 30, 187–191.
- (26) Ullman, J. R. An Algorithm for Subgraph Isomorphism. *J. Assoc. Comput. Machinery* **1976**, 23, 31–42.
- (27) Cerruela García, G.; Luque Ruiz, I.; Gómez-Nieto, M. A. Step-by-Step Calculation of All Maximum Common Substructures Through a Constraint Satisfaction based Algorithm. *J. Chem. Inf. Comput. Sci.* **2004**, 44(1), 30–41.
- (28) Willett, P.; Barnard, J. M.; Downs, G. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 8(6), 983–996.
- (29) Taraviras, S. L.; Ivanciuc, O.; Carbol-Bass, D. Identification of Groupings of Graph Theoretical Molecular Descriptors Using a Hybrid Cluster Analysis Approach. *J. Chem. Inf. Comput. Sci.* **2000**, 40(5), 1128–1146.

CI0342831