CROATICA CHEMICA ACTA
CCACAA **78** (2) 275–281 (2005)

ISSN-0011-1643
CCA-3008
Original Scientific Paper

From Trees (Dendrograms and Consensus Trees) to Topology*

Guillermo Restrepo[a], and José L. Villaveces[b]**

[a]*Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia*

[b]*Observatorio Colombiano de Ciencia y Tecnología, Bogotá, Colombia*

RECEIVED FEBRUARY 16, 2005; REVISED MARCH 10, 2005; ACCEPTED MARCH 15, 2005

Keywords
chemotopology
mathematical chemistry
topology
cluster analysis
dendrograms
consensus trees

We describe a methodology to endow a set of chemical interest with a topology. This procedure starts with the definition of the chemical set as a group of elements plus their neighborhood relationships. A graphical representation of these two conditions is a dendrogram (tree). Next, we show a mathematical procedure to build up a basis for a topology with which we can calculate several topological properties, such as: closures and boundaries of sets of chemical interest. We show four practical examples of this methodology: 72 chemical elements, 31 steroids, 250 benzimidazoles and 20 amino acids.

INTRODUCTION

There are several chemical systems which are characterized for the relationships among their elements, for example: the chemical elements, acids, bases and families of organic compounds, among others. These relationships are interpreted as similarity relationships.[1] It means, two elements of a set of chemical interest are strongly related if they are very similar. Thus, the study of similarity is essential in this respect. One way to quantify the similarity starts with the definition of every chemical object (element, compound) in mathematical terms;[2] normally as a vector of its attributes or properties.[3] In other words, the similarity studies start with the transformation of the chemical object into a mathematical one. Once the chemical object is defined mathematically, then the similarity among all vectors is calculated by means of a similarity function, commonly related to the distance among themselves.[2,3] A methodology that has shown important results trying to find similarities in chemistry is the cluster analysis, which, taking advantage of several grouping methodologies, finally shows clusters of elements that share common features. These groups or clusters can be interpreted as groups of similar objects. Normally, a way to visualize such clusters, independent of the dimension of the space, or the number of features that determine every chemical object, is a two-dimensional graphical representation called a dendrogram or tree, in mathematical terms. Generally, cluster analysis finishes with the obtention of the dendrogram and its respective analysis and interpretation. But as we showed,[3–8] it is possible to interpret a dendrogram and its clusters as a map of neighborhoods of elements and extracting the notion of neighborhood from these clusters in terms of similarity. It means, if an object belongs to a particular cluster, then this one and the rest of the objects in the same cluster are neighbors of it, due to the fact that they are similar by construction.

* Dedicated to Dr. Edward C. Kirby on the ocassion of his 70th birthday.

** Author to whom correspondence should be addressed. (E-mail: grestrepo@unipamplona.edu.co)

Since it is possible to define a neighborhood for every element of the set, we can approach this interpretation and apply the mathematical theory in charge of studying neighborhood relationships, which is the Topology. With this tool it is possible to define topologies within the set and to study some topological properties in the set, as closures and boundaries, among others.[8] We called this methodology »chemotopology« due to the fact that it combines cluster analysis that takes part of chemometrics or chemoinformatics and topology. But, what is a topology? A topology of a set is the set itself and a collection of the neighborhoods of its elements.[9–11] Taking advantage of this definition and knowing the nature of chemical sets, we can say that this sort of uses of general topology can have further application in chemistry since one of the main characteristics of the chemical sets is the relationship among their elements.[12] Recently,[3–7] we showed through this chemotopological procedure that the mathematical boundary of metals in the set of chemical elements is the set of semimetals. Then, taking a topological advantage of a dendrogram, or in general a tree, it is possible to find out some well-known relationships or in other cases new relationships. But, as we mentioned above, there are several chemical systems of interest built up using similarities of their elements. Besides, there are some works, reported in literature, using cluster analysis of some chemical sets, such as: benzimidazoles,[13] amino acids,[14] steroids[15] and chemical elements.[3–7] Our aim in this paper is to show the chemotopological methodology and apply it to the study of those chemical sets.

METHODOLOGY

The general procedure of cluster analysis can be divided in two steps: measurements of similarities and grouping methodologies. The first step includes the selection of one measure of similarity, normally a metric one,[2,16–18] which is applied to calculate the similarity relationships among all chemical objects. The second step of cluster analysis is the selection of a grouping methodology[17,18] that in mathematical terms implies the selection of a way to calculate the distance between one point and a set. The final product of these two steps is a hierarchical classification of the set that can be represented in a graphical way. The most common graphical representation is a dendrogram that shows the clusters obtained through the two steps mentioned above. Normally, cluster analysis studies use only one similarity function and only one grouping methodology to finally obtain only one dendrogram (Figure 1a), but there is an arbitrariness in the choice of a particular similarity function and a grouping methodology; then as we showed recently,[3,4,7] it is recommendable to obtain consensus trees to search for those features common to several of the employed methods (similarity function and grouping methodology). A hypothetical consensus tree appears in Figure 1b. Thus, we can have two different representations of similarity relationships: dendrograms and consensus trees (Figure 1). However, they can be interpreted, in general, as trees (acyclic and connected graphs), and in this way we can talk indistinctively about dendrograms and consensus trees.

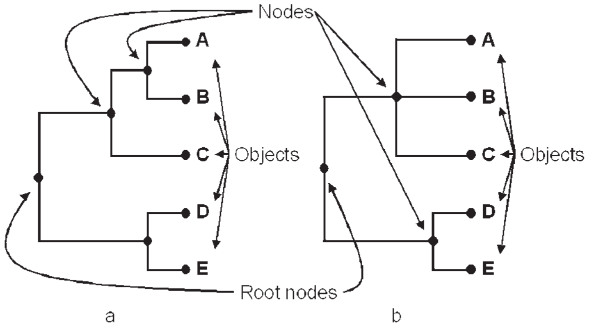

Figure 1. a) A dendrogram; b) A consensus tree and their vertices.

Definition 1. – A tree is a graph showing the clusters of a set of objects, with the following classes of vertices:

1. vertices of degree 1, corresponding to objects;
2. vertices of degree greater than 3, called nodes;
3. only one vertex of degree 2, called root node.

We show in Figure 1 the different vertices in a dendrogram and in a consensus tree.

With the aim of providing the set of chemical interest with a topology we introduce the following definitions (some basic concepts of topology appear in Appendix A1–A2).

Definition 2. – A subgraph G of a tree T is called subtree if:

1. G does not contain the root node;
2. There is a node p of T with degree greater than 1 such that G corresponds to one of the connected subgraphs obtained subtracting p from D.

Definition 3. – Let an n-subtree be a subtree of cardinality less than or equal to n.

Definition 4. – A maximal n-subtree is an n-subtree such that there is no other n-subtree containing it.

We build up a basis for a topology by means of the following theorem, whose proof appears in Appendix (A3).

Theorem 1. – Let Q be a set of chemical interest and $\mathbf{B}_n = \{B \subseteq Q \mid$ be formed by the elements of some maximal n-subtree$\}$. Then, $\mathbf{B}_n$ is a basis for a topology in Q.

Thus, it happens that for every n, different topologies may appear. When $n = 1$ we have a basis which is a collection of all objects (Appendix A4). This means that the neighborhood of every object is itself and not any other; on the other hand, if we have $n = |Q|$, where $|Q|$ is the cardinality or number of elements of the set Q, then the basis is a collection of only one set, the whole set (Appendix A5). For this reason we should select $1<n<|Q|$ and as we showed for the set of 72 chemical elements,[3–7] a selection of $n = 5$ produces a basis which generates a topology that reproduces some of the intuitive ideas about the chemical elements, such as: the classification of metals, non metals and semimetals.

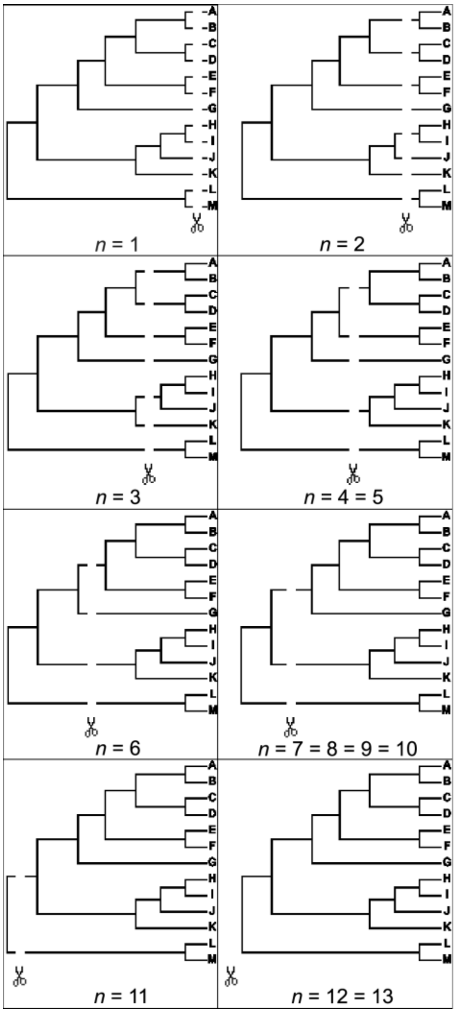

Figure 2. Representation of the influence of *n* in the size of the branches.

The methodology developed to endow a set of chemical objects with a topology is based on the selection of »branches« of the tree to build a basis for the topology. Besides, every choice of *n* determines the size of branches on the tree that we select as neighborhoods of the objects on the tree. In Figure 2, we show a graphic representation of this methodology.

Once we have endowed the set *X* with a topology τ_n we can study some topological properties of the set *X* such as those that appear in Appendix A6.

Now we can apply this methodology to sets of chemical interest.

SOME CHEMICAL EXAMPLES

We show four examples of this methodology applied to chemical systems in the following.

Chemical Elements

Making use of this procedure we build up a topology on the set of 72 chemical elements (Z = 1–86, omitting 58–71) every one defined by 31 physico-chemical properties and found that alkali metals and noble gases are subsets which are not related to other elements.[3–6] It means that in the space of chemical elements these two groups are disjuncts. On the other hand we found that the boundary of metal and non-metals is the same subset of elements, the set of semimetals.

Benzimidazoles

A classification of 238 benzimidazoles making use of graph theoretical and quantum mechanical calculations was made by Niño, Daza, and Tello.[13] In this work the authors developed a dendrogram using Euclidean distance as similarity measure and nearest linkage as grouping methodology. We do not show the dendrogram due to its size, but it may be requested from the authors.

In this set of substances it is possible to classify compounds according to their pharmacological activity; thus we have 5 classes: Angiotensin II (A), Antivirals (AV), Cardiotonics (C) and Antihelmintics (H). Cardinalities of every class were: |A|=158, |AV|=15, |C|=32, |H|=33. Now, taking advantage of the dendrogram shown in that work, we build up a topology τ_{15} on the set of benzimidazoles. Topological properties to these subsets are the following:

$$\overline{\mathrm{A}} = \mathrm{A} \cup \{h_{31}\}$$

$$b(\mathrm{A}) = \{a_{8a}, a_{6a}, a_{18}, a_{12}, a_{16}, a_{13}, a_{15}, a_{5}, a_{11}, a_{4}, a_{3}, a_{14}, a_{10}, h_{31}, a_{0}\}$$

$$\overline{\mathrm{AV}} = \mathrm{AV} \cup \{h_{42}\}$$

$$b(AV) = \begin{Bmatrix} h_{42},\ av_{12},\ av_{10},\ av_{14},\ av_{15},\ av_{11},\ av_{9}, \\ av_{8},\ av_{7},\ av_{5},\ av_{4},\ av_{6},\ av_{3},\ av_{2},\ av_{1} \end{Bmatrix}$$

$$\overline{\mathrm{C}} = \mathrm{C}$$

$$b(\mathrm{C}) = \varnothing$$

$$\overline{\mathrm{H}} = \mathrm{H} \cup \begin{Bmatrix} av_{12}, av_{10}, av_{14}, av_{15}, av_{11}, av_{9}, av_{8}, av_{7}, av_{5}, av_{4}, av_{6}, av_{3}, av_{2}, \\ av_{1}, a_{8a}, a_{6a}, a_{18}, a_{12}, a_{16}, a_{13}, a_{15}, a_{5}, a_{11}, a_{4}, a_{3}, a_{14}, a_{10}, a_{0} \end{Bmatrix}$$

$$b(\mathrm{H}) = \begin{Bmatrix} av_{12}, av_{10}, av_{14}, av_{15}, av_{11}, av_{9}, av_{8}, av_{7}, av_{5}, av_{4}, av_{6}, av_{3}, av_{2}, \\ av_{1}, a_{8a}, a_{6a}, a_{18}, a_{12}, a_{16}, a_{13}, a_{15}, a_{5}, a_{11}, a_{4}, a_{3}, a_{14}, a_{10}, a_{0} \end{Bmatrix}$$

From this we can say that the classification of benzimidazoles in 4 classes does not give 4 disjunct subsets according to the five graph-theoretical and seven quantum descriptors used. We conclude this, due to the presence of some benzimidazoles *h* and *av* in several sets.

But, it is correct to say that the set of cardiotonics is a well-defined group because its closure is itself and its boundary is empty; this result shows that this set is disjunct in the space of benzimidazoles, such as occurs in the set of chemical elements with alkali metals and noble gases.[3–7] On the other hand, according to our results it is possible to speculate and say that those substances that are in the boundary of two subsets can have intermediate properties between two subsets.[8] In this way it would be of special interest to develop studies related to the properties of h_{31}, h_{42} and those elements that belong to b(H).

Steroids

Recently, Bultinck and Carbó-Dorca[15] developed a classification of 31 steroids using molecular quantum similarity and cluster analysis. The chemical structures of this substances appear in Figure 3.

With the aim of studying the results of these authors, we developed a chemical classification of the set of steroids in 5 molecular classes: those able to form the tautomer enol (E); those without multiple bond endocyclic (W); those aromatics (A); those with a doble bond endocyclic (C-5-C-6 of the system cyclopentane-perhydro-phenantrenum) (D) and those conjugated systems not able to form the tautomer enol (C). Once we applied our methodology, we found that to the topology built up for every one of the 3 dendrograms developed by the authors, the topological properties are the same in the dendrogram obtained using the Carbó similarity index and the dendrogram obtained by means of the stochastic transform and Euclidean distance. We show the information of this dendrogram in Figure 4. All five subsets result to be themselves their own closure; *ergo*, their boundaries were empty. These results indicate that this classification of steroids according to chemical knowledge on structure and reactivity gives disjunct sets, or in other words, robust groups.

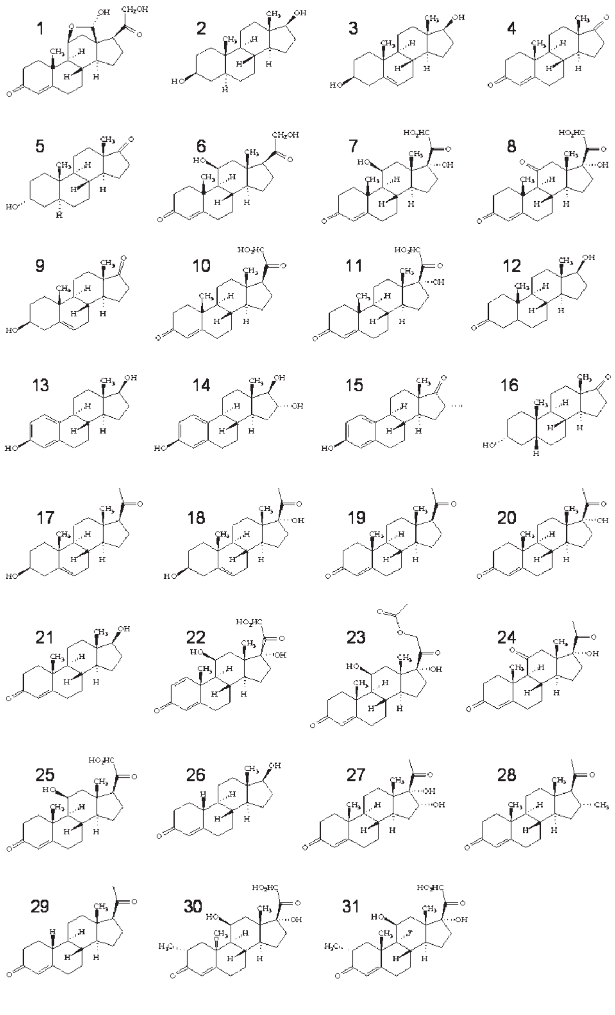

Figure 3. Steroids studied by Carbó-Dorca and Bultinck.

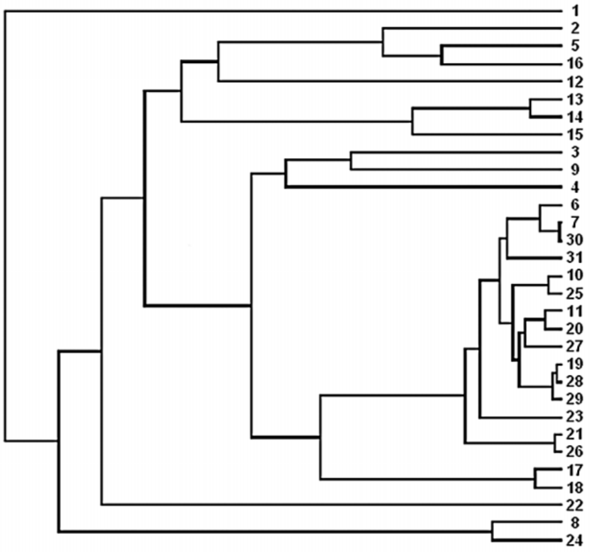

Figure 4. Dendrogram of 31 steroids.

Amino Acids

A recent work related to the behavior of the amino acids within two different environments were reported by Cárdenas *et al.*[14] In this work a set of 20 genetically encoded amino acids (and five of their conformers) were studied with the aim to predict the peptidic properties resulting from the exchange of two amino acids in a proteic chain. The peptidic chain was emulated using two capping models to simulate the effect of its nearest neighbors. These are OCH-X_{aa}-NH_2 and A-X_{aa}-A, where X_{aa} is the conformer of interest and A is alanine. Thus, defining every amino acid system as a vector of 40 *ab initio* quantum chemical and graph-theoretical indices, the authors developed principal component analysis and found 8 principal components. With these results, they performed cluster analysis to every capping model using the more relevant variables suggested by principal component analysis. After this, the authors built up a consensus tree of two dendrograms obtained for every capping model, where the amino acids are represented by their one-letter code plus *h* or *b* that means their alpha or beta backbone conformation respectively, plus g_{+}, g_{-} or *t* for those side chain conformers *gauche*+, *gauche*– and *trans* respectively. The consensus tree cannot be shown in this paper due to its size but may be requested from the authors. Now we build up the basis B_7.

$\mathbf{B}_7$ =

$$\left\{\begin{array}{l}
\{kht, khg_{+}, khg_{-}, mhg_{+}, mhg_{-}, mht\},\\
\{vht, vhg_{-}, vhg_{+}, tht, thg_{+}, thg_{-}\},\\
\{dht, dhg_{+}, dhg_{-}, nht, nhg_{-}, nhg_{+}\},\\
\{qhg_{+}, eht, ehg_{-}, ehg_{+}, qhg_{-}\},\{ah, sht, shg_{-}, shg_{+}\},\\
\{cht, chg_{+}, chg_{-}\},\{iht, ihg_{+}, ihg_{-}\},\{lhg_{-}, lhg_{+}\},\{lht\},\{qht\},\\
\{lbt, lbg_{-}, ibt, ibg_{-}, lbg_{+}, ibg_{+}\},\{vbt, vbg_{+}, vbg_{-}, tbt, tbg_{-}, tbg_{+}\},\\
\{kbt, kbg_{+}, kbg_{-}, mbg_{+}, mbg_{-}, mbt\},\\
\{qbg_{-}, ebg_{-}, ebt, qbg_{+}, ebg_{+}, qbt\},\\
\{sbt, sbg_{+}, ab, sbg_{-}\},\{cbt, cbg_{-}, cbg_{+}\},\{dbg_{+}, dbg_{-}, dbt\},\\
\{nbt, nbg_{-}\},\{ph, pb\},\{fbt, fbg_{-}, fbg_{+}, ybt, ybg_{-}, ybg_{+}\},\\
\{fhg_{+}, fhg_{-}, fht, yhg_{+}, yhg_{-}, yht\},\{hbt, hbg_{+}, hbg_{-}\},\\
\{hhg_{+}, hht, hhg_{-}\},\{wbg_{+}, wbg_{-}, wbt, whg_{+}, wht, whg_{-}\},\\
\{rbg_{+}, rbg_{-}, rbt, rhg_{+}, rhg_{-}, rht\},\{gh, gb\}
\end{array}\right\}$$

In this example we study the common groups of the classification of amino acids which appear in several texts of biochemistry;[19] they are amino acids with hydrophobic side groups (PHO); with hydrophilic side groups (PHI) and those that are in-between (PP). We study these three classes of compounds in the following.

These are the topological properties of the set of amino acids with hydrophobic side groups (PHO):

PHO =

$$\left\{\begin{array}{l}
mhg_{+}, mhg_{-}, mht, vht, vhg_{-}, vhg_{+}, iht, ihg_{+}, ihg_{-}, lhg_{-}, lhg_{+},\\
lht, lbt, lbg_{-}, ibt, ibg_{-}, lbg_{+}, ibg_{+}, vbt, vbg_{+}, vbg_{-}, mbg_{+}, mbg_{-},\\
mbt, fbt, fbg_{-}, fbg_{+}, fhg_{+}, fhg_{-}, fht
\end{array}\right\}$$

$\overline{\text{PHO}}$ =

$$\text{PHO} \cup \left\{\begin{array}{l}
kht, khg_{+}, khg_{-}, tht, thg_{+}, thg_{-}, tbt, tbg_{-}, tbg_{+}, kbt,\\
kbg_{+}, kbg_{-}, ybt, ybg_{-}, ybg_{+}, yhg_{+}, yhg_{-}, yht,
\end{array}\right\}$$

b(PHO) =

$$\left\{\begin{array}{l}
kht, khg_{+}, khg_{-}, mhg_{+}, mhg_{-}, mht, vht, vhg_{-}, vhg_{+}, tht,\\
thg_{+}, thg_{-}, vbt, vbg_{+}, vbg_{-}, tbt, tbg_{-}, tbg_{+}, kbt, kbg_{+}, kbg_{-},\\
mbg_{+}, mbg_{-}, mbt, fbt, fbg_{-}, fbg_{+}, ybt, ybg_{-}, ybg_{+}, fhg_{+},\\
fhg_{-}, fht, yhg_{+}, yhg_{-}, yht
\end{array}\right\}$$

The closure of this set has all lysine on the tree. Thus, lysine is more related to amino acids with hydrophobic side groups that those hydrophilic ones, in spite of its classification as an amino acid with hydrophilic side groups. Besides, in the closure of amino acids of this example all threonine and tyrosine amino acids appears, which are amino acids that are in-between hydrophobic and hydrophilic ones.

Regarding the boundary of this set, we have found that all leucine and isoleucine amino acids appear in the interior of the set and for this reason they do not appear within the boundary. They are the only hydrophobic amino acids that do not appear related to any other amino acid.

In this boundary the only hydrophilic amino acids that appear are those of lysine. Those amino acids that belong to the set of hydrophobic and also appear within the boundary are methionine, valine and phenylalanine. Those amino acids that do not belong to either hydrophobic or hydrophilic are threonine and tyrosine.

Now we show the topological properties of the set of amino acids with hydrophilic side groups (PHI):

PHI =

$$\left\{\begin{array}{l}
kht, khg_{+}, khg_{-}, dht, dhg_{+}, dhg_{-}, nht, nhg_{-}, nhg_{+}, qhg_{+}, eht,\\
ehg_{-}, ehg_{+}, qhg_{-}, qht, kbt, kbg_{+}, kbg_{-}, qbg_{-}, ebq_{-}, ebt, qbg_{+},\\
ebg_{+}, qbt, dbg_{+}, dbg_{-}, dbt, nbt, nbg_{-}, hbt, hbg_{+}, hbg_{-}hhg_{+}, hht,\\
hhg_{-}, rbg_{+}, rbg_{-}, rbt, rhg_{+}, rhg_{-}, rht
\end{array}\right\},$$

$$\overline{\text{PHI}} = \text{PHI} \cup \{mgh_{+}, mgh_{-}, mht, mbg_{+}, mbg_{-}, mbt\},$$

$$b(\text{PHI}) = \left\{\begin{array}{l}
kht, khg_{+}, khg_{-}, mhg_{+}, mhg_{-}, mht,\\
kbt, kbg_{+}, kbg_{-}, mbg_{+}, mbg_{-}, mbt,
\end{array}\right\}.$$

The closure of this set is built up, besides hydrophilic amino acids, of all methionine amino acids, which are hydrophobic amino acids. The boundary shows only lysine and methionine; thus the interior of this set are all amino acids with hydrophilic side groups except lysine.

The last class of amino acids that we studied was the set of those that are in-between amino acids with hydrophobic and hydrophilic side groups (PP).

PP =

$$\left\{\begin{array}{l}
tht, thg_{+}, thg_{-}, ah, sht, shg_{-}, shg_{+}, cht, chg_{+}, chg_{-}, tbt, tbg_{-},\\
tbg_{+}, sbt, sbg_{+}, ab, sbg_{-}, cbt, cbg_{-}, cbg_{+}, ph, pb, ybt, ybg_{-}, ybg_{+}\\
yhg_{+}, ygh_{-}, yht, wbg_{+}, wbg_{-}, wbt, wht, whg_{-}, gh, gb
\end{array}\right\},$$

$$\overline{\text{PP}} = \text{PP} \cup \left\{\begin{array}{l}
vht, vhg_{-}, vhg_{+}, vbt, vbg_{+}, vbg_{-}, fbt,\\
fbg_{-}, fbg_{+}, fhg_{+}, fhg_{-}, fht,
\end{array}\right\},$$

$$b(\text{PP}) = \left\{\begin{array}{l}
vht, vhg_{-}, vhg_{+}, tht, thg_{+}, thg_{-}, vbt, vbg_{+},\\
vbg_{-}, tbt, tbg_{-}, tbg_{+}, fbt, fbg_{-}, fbg_{+}, ybt,\\
ybg_{-}, ybg_{+}, fhg_{+}, fhg_{-}, fht, yhg_{+}, yhg_{-}, yht
\end{array}\right\}.$$

The closure of these amino acids is built up, besides by themselves, by valine and phenylalanine as hydrophobic amino acids. On the other hand, the interior of this set is made up of all amino acids of this class, except threonine and tyrosine which appear as boundary point joined with valine and phenylalanine, both hydrophobic. It is important to remark that this set does not appear related to hydrophilic amino acids.

CONCLUSIONS

The chemotopological methodology applied in this paper shows that given a set of chemical interest (defined by means of its properties) it is possible to apply cluster analysis and topology to evaluate some topological properties of sets of chemical interest, such as: chemical elements, benzimidazoles, steroids and amino acids.

Regarding the chemical elements we found that the mathematical boundary of the set of metals and non-metals is made of semimetals. On the other hand, the results of benzimidazoles show that the classification of them in 4 classes does not give 4 disjunct subsets, due to the fact that there are some benzimidazoles *h* and *av* which appear in several sets. But, it is correct to say that the set of cardiotonics is a well-defined group because its closure is itself and its boundary is empty; this result shows that this set is disjunct in the space of benzimidazoles. Now, we can conclude to the set of steroids that all five subsets result being their own closure themselves, *ergo*, therefore their boundaries are empty. These results indicate that this classification of steroids, according to chemical knowledge on structure and reactivity, gives disjunct sets, or in other words, robust groups. Finally, we can say, regarding amino acids, that the closure of the subset of amino acids with hydrophobic side has all lysines. Thus, lysine is more related to amino acids with hydrophobic side groups that those hydrophilic ones, in spite of its classification as an amino acid with hydrophilic side groups. Besides that, in the closure all threonine and tyrosine amino acids appears, which are amino acids that are in-between hydrophobic and hydrophilic ones. Also, it is possible to conclude that the closure of the subset of amino acids with hydrophilic side groups is built up, plus hydrophilic amino acids, of all methionine amino acids, which are hydrophobic amino acids. The boundary shows only lysine and methionine. The closure of amino acids that are in-between the two above is built up, plus themselves, by valine and phenylalanine as hydrophobic amino acids. It is important to remark that this subset does not appear related to hydrophilic amino acids.

We can say that the chemotopological methodology shown in this paper can be applied not only to the chemical sets shown here but whatever chemical set; in fact, to whatever set, not only of chemical objects but a set in general that can be defined according to the properties or features of its elements.

Acknowledgments. – We thank Dr. C. Cárdenas from Fundación Instituto de Inmunología de Colombia for giving us their respective dendrograms essential to develop this paper. We thank also Dr. J. Rivero from Universidad de Pamplona for drawing the steroids shown in this paper and finally G. Restrepo thanks the Universidad de Pamplona in Colombia and specially Dr. A. González, rector of that University, for their financial support during this research.

APPENDIX

A1. – Let X be a non-empty set and a collection of subsets of X such that:

1) $X \in \tau$

2) $Ø \in \tau$

3) If $O_1,...,O_n \in \tau$, then $\bigcap_{j=1}^{n} O_j \in \tau$

4) If $\alpha \in I$, $O_\alpha \in \tau$, then $\bigcup_{\alpha \in I} O_\alpha \in \tau$

Thus, τ is a topology, the couple (X, τ) is called a topological space and the elements of τ are called open sets.

A2. – Let $\mathbf{B}$ be a collection of subsets of a non-empty set X, such that:

1) $X = \bigcup_{B \in \mathbf{B}} B$

2) If B_1, $B_2 \in \mathbf{B}$, then $B_1 \cap B_2$ is the union of elements of $\mathbf{B}$, then $\mathbf{B}$ is called a basis for the topology τ, where $\tau = \{\bigcup_{B \in F} B \mid F \subseteq \mathbf{B}\}$.

A3. – We should prove that $\mathbf{B}_n$ satisfies the two conditions of Theorem 1.

1) Each object is part of a 1-subtree, of a 2-subtree and so on. Then it is part of a n-subtree. If this subtree is a maximal n-subtree, then the object is already in a maximal n-subtree. If not, then there is a maximal n-subtree that contains it. Thus, each object belongs to some of the maximal n-subtrees and the first condition is satisfied.

2) Any maximal n-subtrees are disjoint. Then, two n-subtrees of the same cardinality are disjoint and if an element belongs to more than one n-subtree of different cardinality, then only that with the highest cardinality is maximal one and for this reason is a member of $\mathbf{B}_n$. Thus, the second condition is satisfied.

A4. – For $n = 1$ we have $\mathbf{B}_1 = \{\{E\} \mid E \in X\}$, and $\tau_1 = P(X)$. This is called the discrete topology, where E is an object and X is a set of objects.

A5. – For $n = |X|$ we have $\mathbf{B}_{|X|} = \{X\}$ and $\tau_{|X|} = \{X, Ø\}$. This is called the indiscrete or coarse topology, where X is a set of objects.

A6. – Some topological properties are the following:

Let $A \subset X$ and $x \in X$; x is said to be a closure point of A if and only if for every $O \in \tau$, such that $x \in O$, then $O \cap A \neq \varnothing$.

Let $A \subset X$; the closure of A is defined as: $\overline{A} = \{x \in X \mid x$ is closure point of $A\}$.

Let and $A \subset X$ and $x \in X$; x is said to be a boundary point of A if and only if for every $O \in \tau$, such that $x \in O$, then $O \cap A \neq \varnothing$ and $O \cap (X - A) \neq \varnothing$.

Let $A \subset X$; the boundary of A is defined as: $b(A) = \{x \in X \mid x$ is boundary point of $A\}$.

REFERENCES

1. D. H. Rouvray, *J. Chem. Inf. Comput. Sci.* **32** (1992) 580–586.
2. P. Willett, J. M. Barnard, and G. M. Downs, *J. Chem. Inf. Comput. Sci*. **38** (1998) 983–996.
3. G. Restrepo, H. Mesa, E. J. Llanos, and J. L. Villaveces, *Topological Study of the Periodic System*, in: D. H. Rouvray and R. B. King (Eds.), *The Mathematics of the Periodic Table*, Nova, New York, In press, chapter 5.
4. G. Restrepo, H. Mesa, E. J. Llanos, and J. L. Villaveces, *J. Chem. Inf. Comput. Sci*. **44** (2004) 68–75.
5. G. Restrepo, H. Mesa, E. J. Llanos, and J. L. Villaveces, *Topological Study of the Chemical Elements*, in: P. Willett (Ed.), *Proceedings of Third Joint Sheffield Conference on Chemoinformatics*, The University of Sheffield, Sheffield, United Kingdom, 2004, Abstract 32.
6. G. Restrepo and J. L. Villaveces, *From Dendrograms to Topology*, in: A. Graovac, B. Pokrić, and V. Smrečki (Ed.), *Proceedings of The 19th Dubrovnik International Course & Conference on the Interfaces among Mathematics, Chemistry and Computer Sciences*, Inter-University Centre, Dubrovnik, Croatia, 2004, p. 68.
7. G. Restrepo, E. J. Llanos, and H. Mesa, *Chemical Elements: A Topological Approach*, in: T. Simos and G. Maroulis (Eds.), *Proceedings of The International Conference of Computational Methods in Sciences and Engineering 2004*, VSP, Athens, Greece, 2004, pp. 753–755.
8. G. Restrepo, H. Mesa, E. J. Llanos, and J. L. Villaveces, *J. Math. Chem*. In press.
9. B. Mendelson, *Introduction to Topology*, Dover, New York, 1990, pp. 1–28.
10. S. Lipschutz, *General Topology*, McGraw-Hill, New York, 1965, pp. 47–86.
11. S. B. Nadler, *La definición de una topología*, Taller de publicaciones de matemáticas de la Facultad de Ciencias UNAM, Ciudad de México, 2002, pp. 125–145.
12. J. Schummer, *HYLE*. **4–2** (1998) 129–162.
13. M. Niño, E. E. Daza, and M. Tello, *J. Chem. Inf. Comput. Sci*. **41** (2001) 495–504.
14. C. Cárdenas, M. Obregón, E. J. Llanos, E. Machado, H. J. Bohórquez, J. L. Villaveces, and M. E. Patarroyo, *Comput. & Chem*. **26** (2002) 667–682.
15. P. Bultinck and R. Carbó-Dorca, *J. Chem. Inf. Comput. Sci*. **43** (2003) 170–177.
16. J. M. Barnard and G. M. Downs, *J. Chem. Inf. Comput. Sci*. **32** (1992) 644–649.
17. R. G. Brereton, *Chemometrics: Applications of Mathematics and Statistics to Laboratory Systems*, Ellis Horwood, Chichester, 1993, pp. 244–262.
18. M. Otto, *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, Wiley-VCH, Weinheim, 1999, pp. 148–156.
19. D. L. Nelson, A. L. Lehninger, and M. M. Cox, *Lehninger Principles of Biochemistry*, W. H. Freeman, New York, 2000, p. 78.

SAŽETAK

Od stabala (dendograma i stabala usaglašavanja) do topologije

Guillermo Restrepo i Josè L. Villaveces

Opisana je metodologija koja topologiju pridružuje kemijski zanimljivim skupovima, i to tako da se definiciji takvih skupova kao grupe elemenata doda još relacija susjedstva, što se grafički opisuje dendogramom (stablom). Dalje je prikazan topološki postupak koji omogućava računanje niza topoloških svojstava, uključivo zatvorenost i granice kemijskih skupova, a koji postupak je onda primjenjen na 72 kemijska elementa, 31 steroida, 250 benzimidazola i 20 amino kiselina.