Journal of Mathematical Chemistry, Vol. 39, No. 2, February 2006 (© 2005)
DOI: 10.1007/s10910-005-9013-5

On the topological sense of chemical sets

Guillermo Restrepo*
Laboratorio de Química Teórica, Universidad de Pamplona, Ciudad Universitaria, Pamplona, Colombia
E-mail: grestrepo@unipamplona.edu.co

Héber Mesa
Departamento de Matemáticas, Universidad del Valle, Edificio 320, Oficina 3190, Cali, Colombia

José L. Villaveces
Grupo de Química Teórica, Universidad Nacional de Colombia, Bogotá, Colombia

Received 29 September 2004; revised 4 November 2004

We describe a mathematical methodology to endow a set of chemical interest with a topology. The procedure starts from a hierarchical classification of the set in a dendrogram (complete binary tree). Then, we cut "branches" of the tree by means of a mathematical procedure and we build up a basis for a topology with these branches. Finally, we show the way to calculate some topological properties, such as; closure, derived set, boundary, interior and exterior of subsets of chemical interest within the particular chemical set chosen at the beginning as object of study.

KEY WORDS: topology, cluster analysis, dendrograms, trees, mathematical chemistry

AMS subject classification: 54A10, 05C05, 80A50

1. Introduction

There are several chemical systems which are characterized by the similarity relationships among their elements, such as; groups of chemical elements, alkanes, ketones, acids, bases, among others. Thus, two elements of a set of chemical interest are "very similar" if they are strongly related; trouble arises when one asks how to quantify such a relationship. One way to put this similarity in a numerical system is to define each element as a point in a mathematical space and to calculate its relationship with other elements by means of a similarity function, commonly a distance function [1–3]. In this procedure every element is defined by means of several features of itself. The number of these features

*Corresponding author.

determines the dimension of the space in which we consider that element as a point. A methodology that has shown important results trying to find similarities among elements is cluster analysis [1,2,4] which, finally shows groups of elements that share common features. These groups or clusters can be interpreted as groups of similar elements. A way to visualize such clusters, independent of the dimension of the space, is a two-dimensional graphic representation called dendrogram [5]. Cluster analysis finishes with the obtention of the dendrogram and its respective analysis and interpretation [1,2]. However, this interpretation, finally, depends on the point of view of the analyst, where some clusters are relevant and others not so much. This fact introduces in this final procedure several undesirable arbitrarities. But as we showed [5–8], it is possible to interpret a dendrogram and its clusters as a map of neighbourhoods of the elements; and extracting from these clusters such similarity neighbourhoods. It means, if an element belongs to a particular cluster, then itself and the rest of the elements of the cluster are neighbours of it since they are similar by construction.

Since it is possible to define a neighbourhood for every element of the set, we can approach this interpretation and apply the mathematical theory in charge of studying neighbourhood relationships, that is Topology [5–8]. With this tool it is possible to define topologies on the set and to study some topological properties of itself: closure, derived set, boundary, interior and exterior. Recently [5–8], we showed, through this procedure, that some of the semimetals belong to the boundary of metals on the set of chemical elements. Then, taking a topological approach of a dendrogram (complete binary tree), it is possible to find out some well-known relationships or in other cases new relationships. With the application of this methodology we can, finally, talk about a mathematical structure of chemical systems [9] and we can do a strict (mathematical) interpretation of a dendrogram where there are no arbitrarities in the interpretation of clusters.

2. Methodology

If we have a chemical set Q of m elements x_i, where every one is defined as a vector $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$ of its n properties, then we can apply cluster analysis to this set with the aim of knowing the similarity relationships among x_is. First, we build up a matrix of elements $(m \times n)$ and calculate by means of a similarity function [1–3,5–8] (frequently a metric [9]) the similarity among all the elements. Thus, we build up a new matrix $(m \times m)$ called similarity matrix [1,4,10] and, using a grouping methodology, we obtain clusters of elements. These clusters are represented in a dendrogram, which is independent of the dimension n of the space of work because it is always two-dimensional. An example of one of these is the one appearing in figure 1.

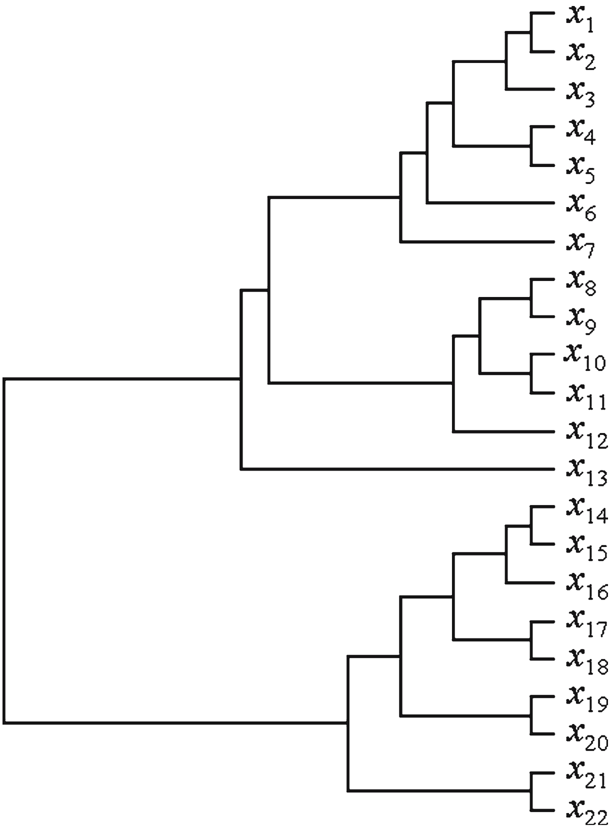

Figure 1. A dendrogram.

Now that we have a dendrogram we can interpret it as a figure which shows relationships among elements, more specifically, neighbourhood relationships. For example, from figure 1 we can say that the neighbourhood of x_1 is itself and x_2, or if we do not wish to be so strict, we can say that x_3 belongs to the neighbourhood, too. However, we do not say that x_1 shares its neighbourhood with x_{14} or x_{22}, for instance. It means that an element belongs to the neighbourhood of another element if they both belong to the same "branch" in the dendrogram [5,11]. The above is an intuitive point of view of the information shown by a dendrogram, but we can put this intuition into mathematical terms taking advantage of the topological sense of a dendrogram [5–8]. We showed a methodology in a recent paper [6] to define these branches as subtrees and we introduced a mathematical procedure to define these subtrees as subgraphs [5]. In the following we show a similar procedure to characterize these branches, but now being based on the codification of every element of the dendrogram.

2.1. Codes on the dendrogram

With the aim of providing the set Q with a topology, we associate a code made of 0s and 1s to every element on the dendrogram. For this purpose we use the system shown in figure 2.

Thus, the dendrogram of figure 1 can be codified and every element can be represented by a code as we show in figure 3.

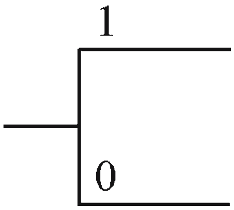

Figure 2. Codification system.

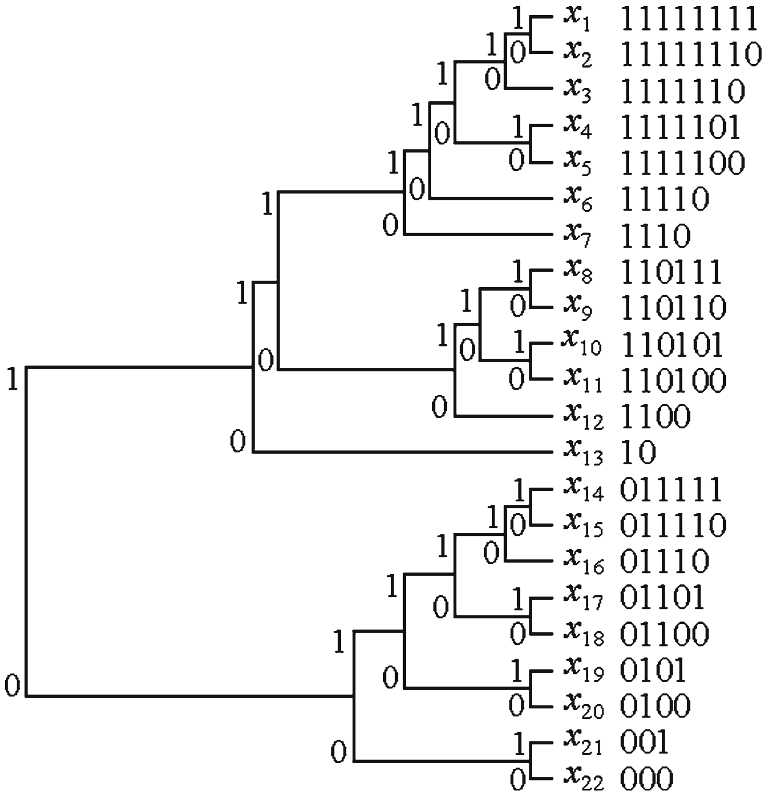

Figure 3. Codes on dendrogram of figure 1.

It is important to remark that the identity of every element is characterized by its code because there is only one code for each element, although the system of codification changes. It means, if we put 0 “above” and 1 “below,” the code of every element changes but its identity remains the same. In spite of its changes of code, there is only one identity for every element.

Now, with these codes we can talk about neighbourhoods in terms of codes. In other words, we can put the intuitive idea of neighbourhood as a branch of the dendrogram in a numerical way. In the following we develop this idea defining subtree in terms of codes.

Definition 1. A subset A of Q is called *subtree* if there is a code $\alpha_1\alpha_2\ldots\alpha_k$ such that:

1. If $x \in A$, then the former components of code of x coincide with α_1 $\alpha_2\ldots\alpha_k$.
2. If for each $y \in Q$ the former components of code of y coincide with α_1 $\alpha_2\ldots\alpha_k$, then $y \in A$.

In other words, a subtree is the set $\{y \in Q |\ y$ starts with the code $\alpha_1\alpha_2\ldots\alpha_k\}$.

Example 1. From the dendrogram in figure 3 we can see that the subset $R = \{x_8, x_9, x_{10}, x_{11}\}$ is a subtree because all its elements start from the code 1101, besides there is no other element in Q such that its code starts from 1101.

But, if we consider the subset NR $= \{x_1, x_2, x_4, x_5\}$, we can see that all these elements have codes that start from 11111, but NR does not have all elements of the tree that start from 11111, it is missing x_3 whose code is 1111110. For this reason, NR is not a subtree.

Now, in order to study neighbourhood relationships, it means, to do a topological study of a set, we define n-subtree.

Definition 2. An *n-subtree* is a subtree of cardinality less than or equal to n.

It means, that an n-subtree has at most n elements. In this way, the subtree R of example 1 is an example of 4-subtree or 5-subtree, or in general l-subtrees where $l \geqslant 4$. For this reason R cannot be either a 3-subtree or a 2-subtree. Now we introduce the definition of maximal n-subtree.

Definition 3. A *maximal n-subtree* is an n-subtree such that there is no other n-subtree containing it.

Once again, taking advantage of example 1 we have that R is a maximal 4-subtree. It is important to say that R, despite being a 5-subtree, is not a maximal 5-subtree since it is contained in $T = \{x_8, x_9, x_{10}, x_{11}, x_{12}\}$ and T is a

5-subtree also because it contains all the elements of Q that have the code 110. On the other hand T is a maximal 5-subtree.

This concept is very relevant to introduce a topology in Q, because every element belongs to one and only one maximal n-subtree. Thus, we identify the neighbourhoods that will build up the topology on Q.

At this point we need some fundamental concepts on the set theory (A1) and topology (A2–A4) to continue introducing our methodology that starts from dendrograms and maximal n-subtrees, and ends by building up a topology on the set Q of chemical interest.

Proposition 1. $\mathfrak{B}_n$ is a partition of Q, where $\mathfrak{B}_n = \{B|B \text{ is a maximal } n\text{-subtree}\}$.

Proof. According to A1, we need to prove two conditions:

1. We will prove that $\bigcup_{B\in\mathfrak{B}_n} B = Q$.
   It is evident that $\bigcup_{B\in\mathfrak{B}_n} B \subseteq Q$, because every $B \subseteq Q$. Now we study $Q \subseteq \bigcup_{B\in\mathfrak{B}_n} B$, in order to prove the equality. Let $x \in Q$, then $\{x\}$ is an n-subtree, to whatever $n \geqslant 1$. If there is a maximal n-subtree B' such that $\{x\} \subseteq B'$, then $x \in B' \subseteq \bigcup_{B\in\mathfrak{B}_n} B$; if there is no such maximal n-subtree, then $\{x\}$ is a maximal n-subtree and, it means, $x \in \{x\} \subseteq \bigcup_{B\in\mathfrak{B}_n} B$.
2. Let B,$D \in \mathfrak{B}_n$, with $B \neq D$, we will show that $B \cap D = \varnothing$.
   Suppose there is an $x \in B \cap D$ and assume that the code of x is $\alpha_1\alpha_2\ldots\alpha_k$. We need to remember that B and D are subtrees and every one of them contains all the elements that start with the same code $b_1b_2\ldots b_t$ and $d_1d_2\ldots d_l$ respectively (we can suppose, without losing the generality, that $t \leqslant l$). Thus, code $\alpha_1\alpha_2\ldots\alpha_k$, starts by $b_1b_2\ldots b_t$ and $d_1d_2\ldots d_l$ simultaneously. It is only possible if $b_i=d_i$ where $i=1, 2, \ldots, t$; this implies that $D \subseteq B$, then D is not a maximal n-subtree, which contradicts our hypothesis. □

Lemma 1. Every partition defined over a subset is basis for a topology.

The immediate consequence of this lemma is the generation of a basis for a topology on Q (A3), it means that $\mathfrak{B}_n$ produces a topology (A2) on Q by means of arbitrary unions among its elements. In the following we define the topology on Q obtained using this basis.

Definition 4. Let $\tau_n = \left\{\bigcup_{B\in\mathcal{F}} B \mid \mathcal{F} \subseteq \mathfrak{B}_n\right\}$ a topology on the set Q.

Proposition 2. The couple (Q, τ_n) is a topological space.

Proof. It is proved by lemma 1 and proposition 1. □

Once we have provided Q with a topology τ_n we can study some topological properties of subsets of Q such as those that appear in A4. These topological properties are sets associated with subsets of Q.

3. On the chemical meaning of topological properties

In a recent work [5–8] we showed that several intuitive chemical ideas can be explained according to our methodology. For example, we found that the mathematical boundary of metals and non-metals is the same set of chemical elements, that is semimetals [5–8]. This result shows that the ancient concept of semimetal as an element whose properties are not from metals nor from non-metals has a mathematical explanation taking advantage of known properties of chemical elements. On the other hand, we showed [8] that, taking advantage of results from Molecular Quantum Similarity [12], the intuitive classification of steroids according to chemical knowledge on structure and reactivity gives a disjoint set which indicates that this classification is correct. We apply the same methodology to sets of aminoacids [13] and benzimidazoles [14] and we found [8] some chemical species belonging to more than one set; something like happens with semimetals regarding to metals and non-metals [5–8]. These results indicate that there are some substances sharing their properties with substances of other sets commonly considered different. Since these results arose from topological properties of sets Q of chemical interest we consider necessary to give a chemical explanation of these mathematical properties. Let us suppose the following example:

Example 2. Let $Q = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z\}$ and suppose we have the following dendrogram to this set (figure 4). Now, if we take $n = 4$, then we look for 4-maximal subtrees on the dendrogram. Thus, we have the following basis for a topology:
$\mathfrak{B}_4 = \{\{a, b, d\}, \{f, m\}, \{c, g, l\}, \{e, n, p\}, \{t, i\}, \{u, h, o\}, \{v, w\}, \{x, y, q, z\}, \{s, j, k\}, \{r\}\}$. Suppose we are interested in determined subset of Q called A, which is: $A = \{a, b, c, d, e, f, g\}$. Then, we can start to calculate the topological properties of this set A.

3.1. Closure of a subset

To calculate the closure of A (figure 5(a)) we need to know the elements of Q which share their neighbourhoods with the elements of A, which are called closure points. Since we have the neighbourhoods in the basis for a topology $\mathfrak{B}_4$, then we search all elements of Q in $\mathfrak{B}_4$ that share their neighbourhoods with elements in A. Thus, as we define at the beginning every element of Q according to its properties, then the meaning of closure is the set of all elements of Q which are similar to the elements of the subset A under study (figure 5(b)).

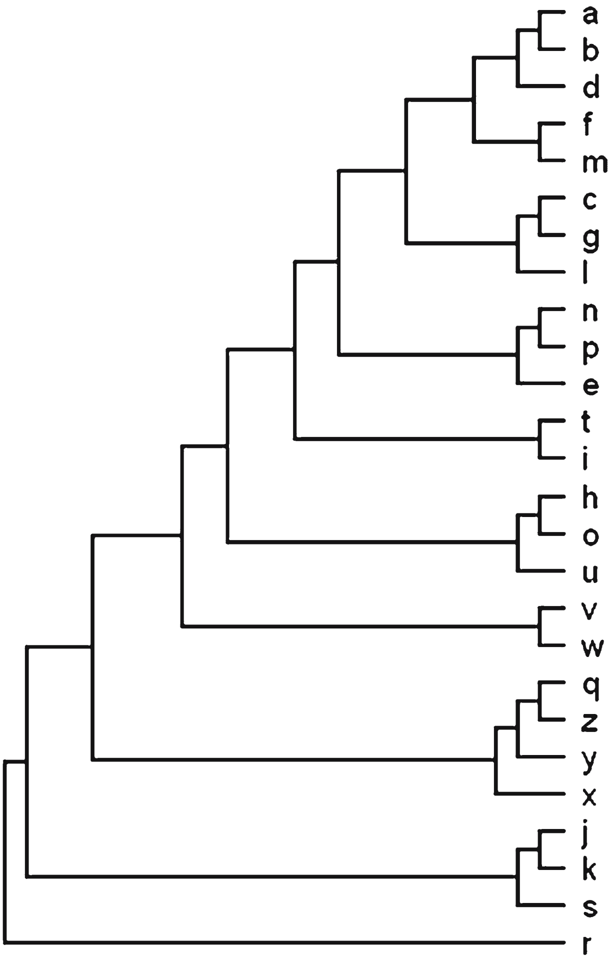

Figure 4. A particular dendrogram of 26 elements.

The above shown mathematically is:

$$\overline{A} = \{a, b, c, d, e, f, g, l, m, n, p\} = A \cup \{l, m, n, p\}.$$

3.2. Derived set of a subset

This set is the one that contains all elements such that if we remove every one of them from their respective neighbourhoods, then these neighbourhoods still remain related to the subset A under study. In other words, they are elements whose neighbourhoods are related to the subset studied A by more than one element. Thus, the meaning of the derived set is the elements of Q similar to the elements in A not only for their similarities but for the similarities of their

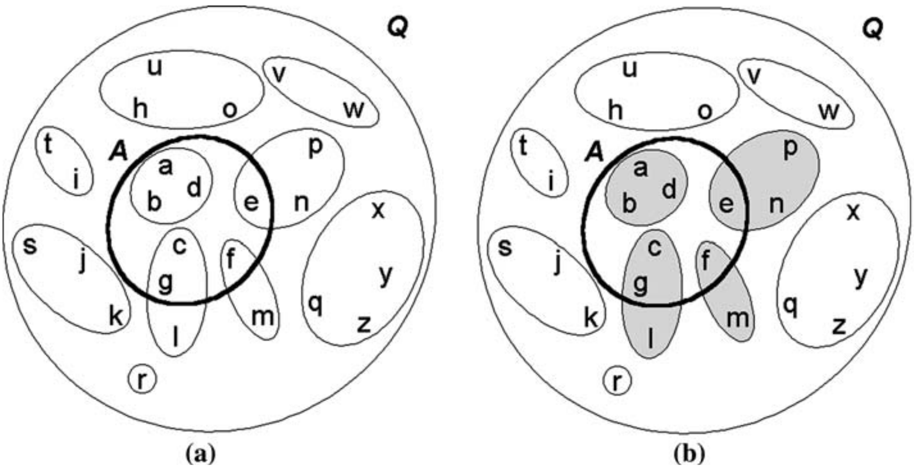

Figure 5. Closure of A according to dendrogram of figure 4.

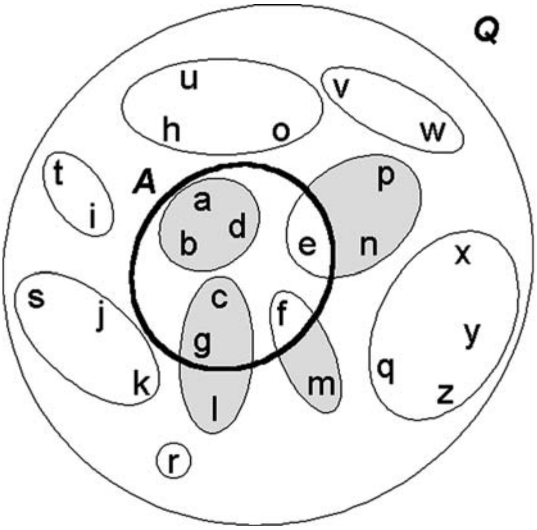

Figure 6. Derived set of A according to dendrogram of figure 4.

neighbours. If a neighbourhood has just one element similar to A, then this element does not belong to the derived set of A since any other of its neighbours is similar to the elements in A (figure 6).

We have in mathematical notation:

$$A' = \{a, b, c, d, g, l, m, n, p\}.$$

3.3. Boundary of a subset

We have in this set those elements of Q whose neighbourhoods have elements either of the A or of the complement of A. Thus, the chemical meaning of this topological property is the elements whose properties are in between properties of the set of interest A and properties from the rest of the elements that do not belong to A (figure 7). Something like happens with semimetals regarding metals and semimetals, for instance.

Which in mathematical terms is:

$$b(A) = \{c, e, f, g, n, p, m, l\}.$$

3.4. Interior of a subset

This is the set of all elements of Q whose neighbourhoods are built up only with elements in A. In chemical terms, we have in this set all elements whose properties are closely related only to the set of interest A (figure 8). If we have an element whose neighbourhood contains elements of A and besides elements of A^c, then this element is not an interior point and it does not belong to the interior set of A.

We have in mathematical terms:

$$\mathrm{Int}(A) = \{a, b, d\}.$$

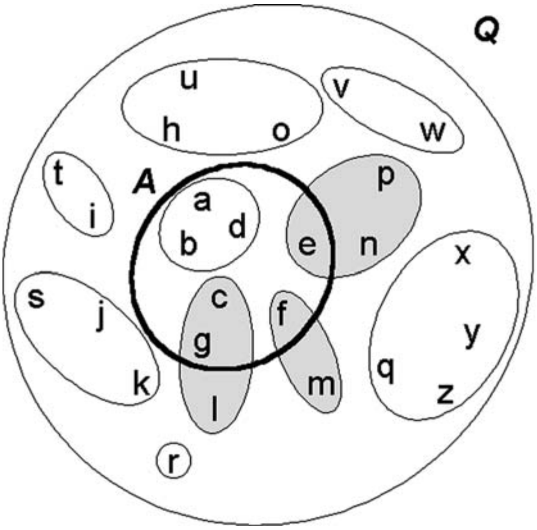

Figure 7. Boundary of A according to dendrogram of figure 4.

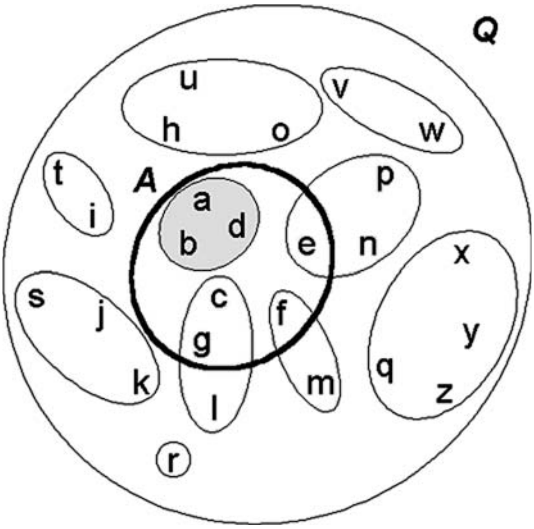

Figure 8. Interior of *A* according to dendrogram of figure 4.

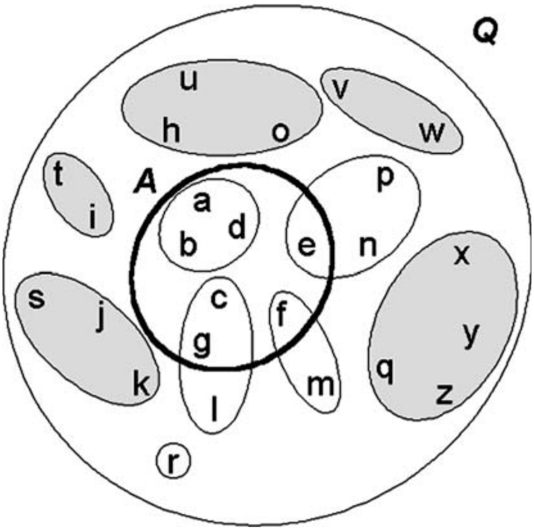

Figure 9. Exterior set of *A* according to dendrogram of figure 4.

3.5. Exterior of a subset

In this set we have all elements of Q such that their neighbourhoods do not contain elements of A. In chemical terms this means, the set of all elements of Q which are not similar to the elements in A (figure 9).

Which in mathematical terms is:

$$\mathrm{Ext}(A) = \{h, i, j, k, o, q, r, s, t, u, v, w, x, y, z\}.$$

4. Conclusions

Chemistry as a science of discrete objects (substances) can be studied by means of discrete mathematics. In this work we developed a mathematical methodology to study sets of chemical objects Q based on a classification of elements in Q. This study starts with a complete binary tree which shows similarity relationships among elements of the set Q. We showed a way to represent every element of the tree as a code, and taking advantage of such codes we defined neighbourhoods on the tree, which represent neighbourhoods on the space of work where all elements of Q are defined. Such neighbourhoods are subsets of Q and they are, besides, a partition of the tree since it is impossible that one element of Q belongs to more than one branch on the tree. We proved that this partition is a basis for a topology which in other words offers the possibility of providing the set Q of interest with a topology. Besides, we described a procedure to calculate some topological properties of subsets of Q, such as closure, derived set, boundary, interior and exterior. Finally, we showed the chemical meaning of such properties in the case of each element of Q has been defined according to its features or properties. In spite of having developed and applied this methodology to chemical systems it is possible to apply the same procedure to other systems such as biological, physical and other ones. The only requirement to apply this methodology is to have a set of discrete elements that can be classified and shown as a tree, and there are many of them in the sciences.

Appendix

A1. Let X be a set non-empty and $\mathcal{P}$ a collection of subsets of X. $\mathcal{P}$ is called a *partition* of X iff:

1. $X = \bigcup_{B \in \mathcal{P}} B$
2. If B_1 and $B_2 \in \mathcal{P}$, then $B_1 \cap B_2 = \varnothing$

A2. Let X be a non-empty set and τ a collection of subsets of X such that:

1. $X \in \tau$
2. $\varnothing \in \tau$
3. If $O_1, \dots, O_n \in \tau$, 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*.

A3. Let $\mathfrak{B}$ be a collection of subsets of a non-empty set X, such that:

1. $X = \bigcup_{B\in\mathfrak{B}} B$
2. If $B_1, B_2 \in \mathfrak{B}$, then $B_1 \cap B_2$ is the union of elements of $\mathfrak{B}$, then $\mathfrak{B}$ is called a *basis for the topology* τ, where $\tau = \left\{\bigcup_{B\in\mathcal{F}} B \mid \mathcal{F} \subseteq \mathfrak{B}\right\}$.

A4. Some topological properties are the following:

Let $A \subset X$ and $x \in X$; x is said to be a *closure point* of A iff 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: $\bar{A} = \{x \in X \mid x$ is closure point of $A\}$.

Let $A \subset X$ and $x \in X$; it is said that x is an *accumulation point* of A iff for every $O \in \tau$, such that $x \in O$, then $(O - \{x\}) \cap A \neq \varnothing$.

Let $A \subset X$; the *derived set* of A is defined as: $A' = \{x \in X \mid x$ is accumulation point of $A\}$.

Let $A \subset X$ and $x \in X$; is said that x is a *boundary point* of A iff 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:

$$b(A) = \left\{x \in X \mid x \text{ is boundary point of } A\right\}.$$

Let $A \subset X$ and $x \in X$; is said that x is an *interior point* of A iff for every $O \in \tau$, such that $x \in O$, then $O \cap (X - A) = \varnothing$.

Let $A \subset X$; the *interior* of A is defined as:

$$\mathrm{Int}(A) = \left\{x \in X \mid x \text{ is interior point of } A\right\}.$$

Let $A \subset X$ and $x \in X$; is said that x is an *exterior point* of A iff for every $O \in \tau$, such that $x \in O$, then $O \cap A = \varnothing$.

Let $A \subset X$; the *exterior* of A is defined as:

$$\mathrm{Int}(A) = \left\{x \in X \mid x \text{ is exterior point of } A\right\}.$$

Acknowledgments

We thank Dr. R. Isaacs from the Universidad Industrial de Santander for his important comments, Dr. Excmo. E. J. Llanos from Observatorio Colombiano de Ciencia y Tecnología for having developed and programmed several algorithms that apply the methodology shown in this paper. Finally, we thank Dr. D. Klein from the University of Texas in Galveston for his contribution to one important definition of this work.

References

[1] R.R. Sokal and P.H.A. Sneath, *Principles of Numerical Taxonomy* (Freeman, San Francisco, 1963).

[2] J.M. Barnard and G.M. Downs, J. Chem. Inf. Comput. Sci. 32 (1992) 644.

[3] P. Willet, J. Chem. Inf. Comput. Sci. 38 (1998) 983.

[4] M. Otto, *Chemometrics, Statistics and Computer Application in Analytical Chemistry* (Wiley-VCH, Weinheim, 1999).

[5] G. Restrepo, H. Mesa, E.J. Llanos and J.L. Villaveces, in: *The Mathematics of the Periodic Table*, eds. R.B. King and D. Rouvray (Nova, New York, 2004, in press), chapter 5.

[6] G. Restrepo, H. Mesa, E.J. Llanos and J.L. Villaveces, J. Chem. Inf. Comput. Sci. 44 (2004) 68.

[7] G. Restrepo, H. Mesa, E.J. Llanos and J.L. Villaveces, *Proceedings of the Third Joint Sheffield Conference on Chemoinformatics* (The University of Sheffield, Shefiield, 2004).

[8] G. Restrepo and J.L. Villaveces, *Proceedings of the Nineteenth International Course & Conference on the Interfaces among Mathematics, Chemistry & Computer Sciences* (Inter-University Center, Dubrovnik, 2004).

[9] B. Mendelson, *Introduction to Topology* (Dover, New York, 1990).

[10] J.V. Crisci, *Introducción a la teoría y práctica de la taxonomía numérica* (OEA, Washington, 1983).

[11] H. Mesa and G. Restrepo, *Proceedings of the Primer Encuentro Nacional de Químicos Teóricos* (Universidad de Pamplona, Pamplona, 2004).

[12] P. Bultinck and R. Carbó-Dorca, J. Chem. Inf. Comput. Sci. 43 (2003) 170.

[13] C. Cárdenas, M. Obregón, E.J. Llanos, E. Machado, H.J. Bohórquez, J.L. Villaveces and M.E. Patarroyo, Comp. Chem. C. 26 (2002) 667.

[14] M. Niño, E.E. Daza and M. Tello, J. Chem. Inf. Comput. Sci. 41 (2001) 495.