MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **60** (2008) 371-384

ISSN 0340 - 6253

On dendrograms and topologies

Héber Mesa[a1], Guillermo Restrepo[b]

[a]Departamento de Matemáticas, Universidad del Valle, Cali, Colombia
[b]Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia

(Received February 20, 2008)

Abstract

We recently developed a methodology to endow a finite set Q with topologies using similarity results from cluster analysis (dendrograms). In this paper we characterise the family of these topologies. We introduce a new method generalising the previous one and allowing to build new topologies over Q not belonging to the former family. Either procedures ensure the existence of a topology given a dendrogram and it is shown that given a topology for Q, mirroring similarities, then a dendrogram can be associated.

1 Introduction

Normally, in classification processes, namely cluster analysis, once the classes are found the study is addressed to the elements of each class, therefore the classes are individually studied. Consequently, relevant information pertaining to similarity among classes is neglected. A method solving this drawback is the chemotopological one [1, 2, 3], which permits to analyse the similarities of any class based upon the similarities found by clustering. In this method, knowledge on the similarity among elements of a set Q is used to draw a complete similarities landscape of any subset of Q.

Chemotopology was originally developed to show the important role of similarities for the trends found in the periodic table of the chemical elements, for example to show that the boundary of the non-metals is the set of semimetals [1, 3, 4]. However, it has found application in the study of other chemical sets, e.g. amino acids, benzimidazoles, steroids [5] and forth row monohydrides [6]. Although chemotopology has been applied in chemistry, it is not restricted to this science; in fact, chemotopology is a general mathematical method able to deal with any set whose elements are defined by their features.

[1]E-mail: hebermes@univalle.edu.co

In reference [2] we showed how different topological ideas, regarding the chemical set under study, can be used to derive topological conclusions with chemical sense. These ideas result from the calculation of topological properties by using chemotopology. Some of these properties are closure, boundary and interior. In short, given a set Q the closure of $A \subset Q$ contains the elements of Q similar to A; the elements of Q similar to A and simultaneously to elements not included in A constitute the boundary of A. The interior of A contains the elements of Q which are completely similar to A and constitute the "core" of A. A further mathematical discussion on the topological properties and on their chemical meaning is given in reference [2]. An important aspect of chemotopology is the generalised concept of similarity that can be derived by its application. Chemotopology permits to reach a deep understanding of the similarity relationships among members of a set, it permits to find elements of a class which are strongly related to the main features of the class, i.e. class representatives; additionally it is possible to find elements which share features of different classes and therefore are transition elements between different classes. All these issues are common in chemical research, for example in drug design it is always wanted to know the nearness or similarity of different pharmacophores in order to save time and resources in developing new medicines. Historical examples of this kind of thought abound, e.g. the similarities among chemical elements studied by Mendeleev which lead to the periodic table; or the development of the transistor by spiking a material with a similar one. More recently, Stadler and co-worwers have brought interesting ideas on the use of topology, closures, boundaries, connectedness, convergence and continuity in fields like combinatorial chemistry and genotype [7, 8, 9, 10, 11]. These, and many other examples are deeply rooted in the idea of nearness or similarity. That notion of nearness is the workhorse of topology and that is what chemotopology studies.

In this paper we explore the mathematical foundations of chemotopology and we characterise the family of topologies for a set Q that are obtained by the application of the method. It is described the relationship between the cardinality of open sets, i.e. similarity neighbourhoods, and an integer number. Finally, we generalise the chemotopological method as the procedure where all similarity neighbourhoods of any element in Q are regarded as open sets of the topological basis and they are not restricted to an integer. The family of topologies obtained by this generalised method is characterised and its relation with the former family of topologies is studied.

2 Usual Chemotopological method and its topologies

We call usual chemotopological method the one depending on an integer number [1, 2, 3, 4, 5, 6]. Since chemotopology uses cluster analysis results, namely a dendrogram, to look for topologies under a set Q, we show a hypothetical dendrogram D (Figure 1) defined

on Q. It is noteworthy that a dendrogram is obtained through the common process of cluster analysis that provides a hierarchical classification of the elements in Q [12, 13]. In order to describe the chemotopological procedure we introduce the following definitions [1, 2, 3]:

Definition 1. A **dendrogram** is a rooted acyclic-binary graph with the following kinds of vertices:

1. Vertices of degree 1, called objects.
2. Vertices of degree 3, called nodes.
3. Only one vertex of degree 2, called root node.

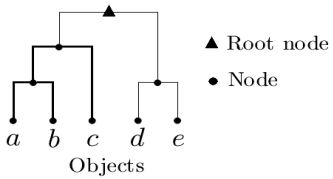

Figure 1: A hypothetical dendrogram of five objects and its types of nodes. Bold lines correspond to a subtree of the dendrogram.

Definition 2. Let G be a subgraph of the dendrogram D. We say that G is a **subtree** iff $G = D$ or:

1. G does not contain the root node (Figure 1) of D, and
2. There exists a node p in D of degree greater than 1 such that G corresponds to one of the connected subgraphs obtained by deleting p from D.

Although subtree is defined as a graph (Definition 2), it is also associated with a subset of Q (where Q is the set of nodes of degree 1 in D) made from all the elements that are nodes of degree 1 in G. Hence, when we refer to the cardinality of G, we mean the cardinality of this associated set. For instance, $G = \{a, b, c\}$ is the set associated to the subtree containing the elements a, b and c in Figure 1 (bold lines).

Definition 3. Let G be a subtree and n a positive integer. We say that G is a ***n*-subtree** iff $|G| \leq n$.

Thus, a n-subtree can be a subtree with less than n objects; then, for every $k \leq n$ we have that every k-subtree is simultaneously a n-subtree. In particular, a 1-subtree can be regarded as a n-subtree for every n.

Definition 4. Let G be a n-subtree. We say that G is a **maximal *n*-subtree** iff it is not contained in any other n-subtree.

Definition 5. Let $\mathfrak{B}_n$ be the family of subsets of Q holding:

$$\mathfrak{B}_n = \{G \subseteq Q \mid G \text{ is a maximal } n\text{-subtree}\}.$$

Theorem 1. Let Q be a non-empty finite set, D a dendrogram defined on Q and n a positive integer. Then, (Q, τ_n) is a topological space, where $\tau_n = \left\{ \bigcup_{B \in \mathcal{F}} B \mid \mathcal{F} \subseteq \mathfrak{B}_n \right\}$. τ_n is called a topology obtained through maximal n-subtrees.

The proof of this theorem appears in reference [2] and is based upon the fact that $\mathfrak{B}_n$ is a partition of Q, which guaranties that $\mathfrak{B}_n$ is a basis for a topology. In terms of similarity, the elements in $\mathfrak{B}_n$ are the similarity neighbourhoods of the elements in Q since they come from the branches (n-subtrees) of the dendrogram.

3 Characterising usual topologies

We describe in this section the common feature of the topologies obtained through the usual chemotopological method mentioned in the previous section [1, 2, 3, 4, 5, 6]. The family $\mathfrak{B}_n$ of maximal n-subtrees is a partition of Q [2], which is a consequence of the "maximality" of the subtrees. Each element $x \in Q$ also belongs to an 1-subtree that is simultaneously a n-subtree; *ergo* it is in, at least, one maximal n-subtree. Given two maximal n-subtrees with common elements, it can be proved that one of them ought to be contained in the other one, for this reason they ought to be equal. It can be seen that this fact is enough to obtain the topology of Theorem 1 but it produces strict consequences in the generated topologies. In order to study these consequences, we characterise the topologies found through the usual chemotopological method.

Proposition 1. Let Q be a non-empty finite set, n a positive integer and D a dendrogram on Q. Then, any open set is simultaneously a closed set in the topological space (Q, τ_n).

Proof. Since any open set is the finite union of elements of $\mathfrak{B}_n$ (basic open sets), it is enough to prove that these sets are closed because the finite union of closed sets is closed. Let $O \in \mathfrak{B}_n$, owing to $\mathfrak{B}_n$ is a partition of Q, thereby $O^C = \cup\{B \in \mathfrak{B}_n \mid B \neq O\}$; then $O^C \in \tau_n$, it means O^C is an open set and for this reason O is closed. □

Proposition 1 guarantees that the topology obtained through maximal n-subtrees belongs to a particular class of topologies. This result is independent of the methodology used to calculate the dendrogram and it is also independent of the selection of the n number. This general result is an evidence of the underlying mathematical structure [14] of the research method, therefore of the set Q. An example of a mathematical structure for a set Q is, for instance, the case of the chemical elements [1, 3, 4] where their similarity neighbourhoods endow Q with a topology. In this case the mathematical structure is a topological one.

Definition 6. Let Q be a non-empty finite set. We define $TOP(Q)$ as the set of all topologies defined over Q and $TOP_{OC}(Q)$ as the set of all topologies defined over Q where each open set is simultaneously a closed set.

According to Proposition 1 and Definition 6, given a dendrogram D and an integer n, then $\tau_n \in TOP_{OC}(Q)$ as is shown in Figure 2.

The common process in the usual chemotopological study is: given a dendrogram, to extract its maximal n-subtrees by the selection of a n number, and build a topological basis (Theorem 1). We showed (Proposition 1) that every element of a topology obtained from this basis is an open-closed set of Q. It means that this topology belongs to the family of topologies $TOP_{OC}(Q)$. This allows us to formulate the following question: if we consider a topology τ in $TOP_{OC}(Q)$, there will exist a dendrogram D and an integer n such that $\tau = \tau_n$? This question is equivalent to the following two questions:

1. If we only consider topologies obtained by dendrograms and integers n using maximal subtrees, is it possible to cover the whole family $TOP_{OC}(Q)$?

2. If we consider all the possible dendrograms over Q[2] and the topologies τ_n with all the different integers n, can we obtain all the topologies belonging to $TOP_{OC}(Q)$?

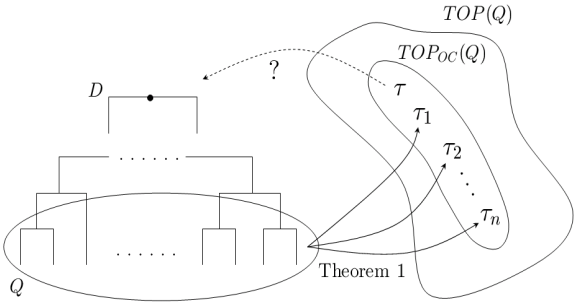

Figure 2: Family of topologies $\tau_n \in TOP_{OC}(Q)$ that can be obtained from a dendrogram D defined on Q through Theorem 1 using an integer number. The dashed arrow between $TOP_{OC}(Q)$ and D rises the question on the possibility of obtaining a dendrogram from a topology τ_n.

In order to answer these questions we developed the concept of O_x, the Lemma 1 and the Proposition 2 (see below). We use the notation O_x to represent the smallest open set containing the element $x \in Q$ in a topology τ. Thus, $O_x = \cap\{O \in \tau \mid x \in O\}$, that is O_x can be obtained by the intersection of all the open sets of the topology τ containing x. Since Q is finite in our case, then O_x is an open set of τ.

[2]The total number of dendrograms $|F|$ that can be defined over a set Q of cardinality N is given by $|F| = \frac{(2N-3)!}{2^{N-2}(N-2)!}$, [15].

Lemma 1. Let Q be a non-empty finite set and $\tau \in TOP_{OC}(Q)$ a topology. Then, the family $\mathcal{B} = \{O_x \mid x \in Q\}$ is a partition of Q.

Proof. If $O_x \neq O_y$ then these sets must have no common elements since if there were a $z \in O_x \cap O_y$, then $O_z \subseteq O_x$. Suppose that $x \notin O_z$, then $x \in O_z^C$ that is an open set, then $O_x \subseteq O_z^C$, which implies that $z \in O_z^C$ and it comes to a contradiction. Then, we must have that $x \in O_z$, which implies that $O_x \subseteq O_z$ and $O_z = O_x$, in the same way $O_z = O_y$ and in conclusion $O_x = O_y$. This result contradicts our first hypothesis. In other words, we have that $\mathcal{B}$ is a partition of Q. □

Proposition 2. Let Q be a non-empty finite set and $\tau \in TOP_{OC}(Q)$ a topology. Then, there exists a positive integer n and a dendrogram D defined on Q such that $\tau_n = \tau$.

Proof. According to Lemma 1 we know that $\mathcal{B} = \{O_x \mid x \in Q\}$ is a partition of Q. On the other hand, $\mathcal{B}$ is basis for the topology τ. Suppose that $B = \{B_1, B_2, \ldots, B_m\}$. We can consider without loss of generality that $|B_1| \geq |B_2| \geq \cdots \geq |B_m|$. Additionally, suppose the dendrograms $D_1, D_2, \ldots, D_m$ defined on the sets $B_1, B_2, \ldots, B_m$, respectively. If we join D_1 and D_2 together by their root nodes to a new root node as shown in Figure 3 and we connect to this new dendrogram the dendrogram D_3 and the process is repeated until linking the dendrogram D_m, then we obtain a dendrogram D defined over $B_1 \cup B_2 \cup \cdots \cup B_m = Q$ (Figure 4).

If we consider $n = |B_1|$ then $\mathfrak{B}_n = \mathcal{B}$, where $\mathfrak{B}_n$ is the set of maximal n-subtrees of D; and for this reason $\tau_n = \tau$. We illustrate this in Figure 4. □

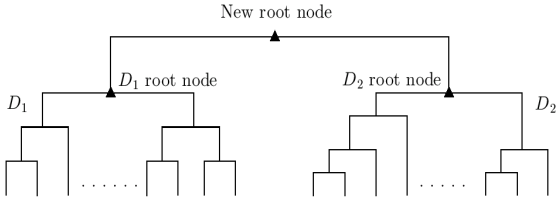

Figure 3: Joining two dendrograms together to build a new one.

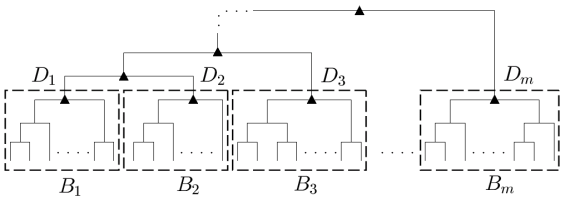

Figure 4: Building the dendrogam D according to the proof of Proposition 2.

Proposition 2 shows that all the possible topologies that can be obtained through

dendrograms and maximal n-subtrees are all the topologies whose open sets are simultaneously closed sets. The most important fact is that, if for any other method (not necessarily using dendrograms) a topology is obtained in such a way that it belongs to the class $TOP_{OC}(Q)$, then that topology has associated a dendrogram and its corresponding n-subtrees. In order to establish a contrast between the usual chemotopological method (maximal n-subtrees) and the one formulated in Proposition 2, we may summarise the usual chemotopological method in the following order:

1. Building a dendrogram on Q.
2. Partitioning Q using maximal n-subtrees.
3. Obtaining a topology τ_n for Q.

In Proposition 2 we raised a reformulation of the order established in the usual chemotopological method [1, 2, 3, 4, 5, 6] in such a way that we can use any order of application of the three steps mentioned above as we show in Figure 5. It means that:

1. If we partition Q by any method, then we can build up a topology for Q and associate a dendrogram to Q.
2. Or we can obtain a topology $TOP_{OC}(Q)$ and, starting from it, obtain a partition on Q and then associate a dendrogram to Q.

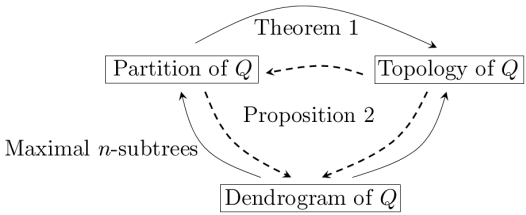

Figure 5: Methodology of the usual chemotopological study (continuous arrows) and new insights raised by Proposition 2 (dashed arrows).

We have changed the direction of the chemotopological process **from dendrograms to topologies**, to, **from topologies to dendrograms**.

4 A new vision

Cluster analysis is a mathematical tool used in several fields of science to find similarities among objects of a set Q [12, 13]. In this way, dendrograms D defined on Q, and their

subtrees, show those similarity relationships. The degree of resemblance among the elements in Q depends on the size of the subtrees we select. In a metaphorical language, the size of the subtrees is like the use of a "magnifying glass" to observe a dendrogram. If we move the magnifying glass closer to the dendrogram, subtrees of low cardinality result; in contrast, if the magnifying glass is moved away, then large subtrees are obtained. An important fact of the way to adjust the magnifying glass or the way to select the maximal n-subtrees is that we cannot "break" subtrees and re-group them to our will. For instance, if we extract the maximal 2-subtrees from the dendrogram depicted in Figure 1, we obtain: $\{a, b\}$, $\{c\}$, $\{d, e\}$. However, we cannot build a topology with the sets $\{a, b, c, d\}$, $\{e\}$, because it does not show the actual similarities among a, b, c, d and e represented in the dendrogram. To justify the impossibility of breaking subtrees and re-grouping them, it is important to note that a topology is constructed using some subsets of Q. The union of these subsets ought to cover Q. Initially, the choice of those subsets is arbitrary, but if we wish to build the topology from a dendrogram D, then we need to be more selective when choosing subsets of Q. One of the attention points for selecting those subsets is to consider the subsets associated to the subtrees of D. The reason for selecting these subsets as the basis for a topology on Q is the fact that we want to build a topology on Q using similarity information. Hence, it is apropriate to build that topology considering those "pieces" containing information about similarities in Q as elements of the basis. These "pieces" are the mentioned subtrees. For instance, we cannot permit that the couple $\{x, y\}$ appears in the basis of the topology if in the dendrogram these two elements are not "directly" joined together (forming a 2-subtree). The reason for this is that in the topology x and y would be inseparable, which means that they cannot be separated by open sets. In a case of this sort, the topology would not correctly represent the dendrogram.

The important fact in the representation of similarities using dendrograms is not the use of maximal n-subtrees but something more general, the concept of subtree. A maximal n-subtree is a particular case of a subtree. It is in the concept of subtree where the similarities underlie [16].

Our aim is to build topologies using subsets of Q, but these subsets ought to represent similarities.

We propose the criterion of the subtrees of the dendrogram as a rule for selecting the members of the basis. In other words, **we propose that the elements of the basis for a topology are only subtrees of the dendrogram *D*.**

Finding a criterion for choosing these subtrees is an interesting discussion. For the special case of topologies τ_n, obtained through maximal n-subtrees, we refer to the selection of the number n. However, in the general case we are describing in this paper, the discussion may be more complicated since, as we show in reference [16], a dendrogram D has exactly $2|Q| - 1$ subtrees. Which of them should we select? Should we consider

the cardinality of subtrees? Should these subtrees be disjoint? These questions and some others are more complex than the selection of a number between 1 and $|Q|$, which is the case when using maximal n-subtrees.

In other words, our aim is to propose a method for selecting the members of the basis not necessarily related to a positive integer n restricting the size of the branches to a particular cardinality. This new vision or new rule for selecting members of the basis considers just the concept of subtree and not the one of maximal n-subtree. In this case we have topologies that can be out of the class $TOP_{OC}(Q)$.

Proposition 3. Let Q be a non-empty finite set, D a dendrogram defined on Q and $\mathfrak{B}_D$ a family of subtrees of D such that $\bigcup_{B\in\mathfrak{B}_D} B = Q$. Then, $(Q, T_{\mathfrak{B}_D})$ is a topological space where $T_{\mathfrak{B}_D} = \left\{ \bigcup_{B\in\mathcal{F}} B \mid \mathcal{F} \subseteq \mathfrak{B}_D \right\}$.

Proof. It is enough to prove that $\mathfrak{B}_D$ is basis for a topology. This is met because if the intersection of two subtrees is non-empty, then one of them should be contained into the other one [2]. Thus, $\mathfrak{B}_D$ is closed under intersections. □

We show a graphical explanation regarding Proposition 3 in Figure 6.

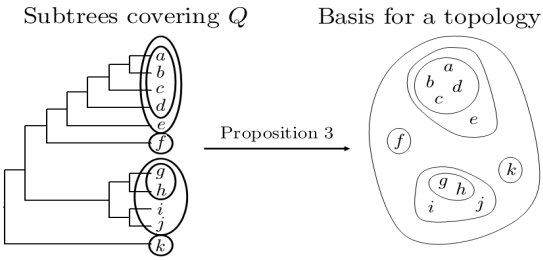

Figure 6: New methodology to build up a basis for a topology.

This proposition offers a new way for generating topologies using dendrograms. In this case, any topology τ_n (Theorem 1) can be seen as a topology obtained by this method; however, the contrary is not true. It means that the procedure generating topologies from maximal n-subtrees is contained in this new procedure based on subtrees. In other words, the new method is more general than the former one based upon maximal n-subtrees; an example of this generality is the following:

We show in Figure 7 a dendrogram D and a family $\mathfrak{B}_D = \{\{a, b\}, \{c\}\}$ of subtrees covering Q. The topology obtained through this family is $T_{\mathfrak{B}_D} = \{\varnothing, \{a, b\}, \{a, b, c\}\}$. This topology does not belong to $TOP_{OC}(Q)$ since the open set $\{a, b\}$ is not a closed set and according to Proposition 2 no dendrogram defined over Q can generate this topology with maximal n-subtrees.

There will exist a specific class of topologies containing all possible topologies that can be obtained through this new method? It means, there will exist a characterisation similar

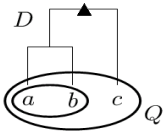

Figure 7: A new kind of topology for Q.

to the one of the topologies τ_n? To answer these questions we developed the following proposition:

Proposition 4. Let Q be a non-empty finite set, D a dendrogram on Q and $T_{\mathfrak{B}_D}$ a topology obtained according to Proposition 3. Then, for all $x, y \in Q$ we have either $O_x \cap O_y = \varnothing$, $O_x \subseteq O_y$ or $O_y \subseteq O_x$.

Proof. Since $\mathfrak{B}_D$ is basis for $T_{\mathfrak{B}_D}$, then O_x is the intersection of all the elements of $\mathfrak{B}_D$ containing the element x, but $\mathfrak{B}_D$ is closed under finite intersections, for this reason $O_x \in \mathfrak{B}_D$. It means that O_x is a subtree of the dendrogram D. In the same way O_y is a subtree of D. On the other hand, we know that if the intersection of two subtrees is non-empty then one of them should be contained into the other one. It means, if $O_x \cap O_y \neq \varnothing$ then either $O_x \subseteq O_y$ or $O_y \subseteq O_x$. □

Definition 7. Let Q be a non-empty finite set. We say that $TOP_P(Q)$ is the family of all topologies $\tau \in TOP(Q)$ such that for every $x, y \in Q$ we have either $O_x \cap O_y = \varnothing$, $O_x \subseteq O_y$ or $O_y \subseteq O_x$.

According to Proposition 4 it is possible to ask similar questions to those of section 2, that is: if we consider a topology τ in $TOP_P(Q)$, there will exist a dendrogram D and a family $\mathfrak{B}_D$ of subtrees of D yielding this topology? (Figure 8).

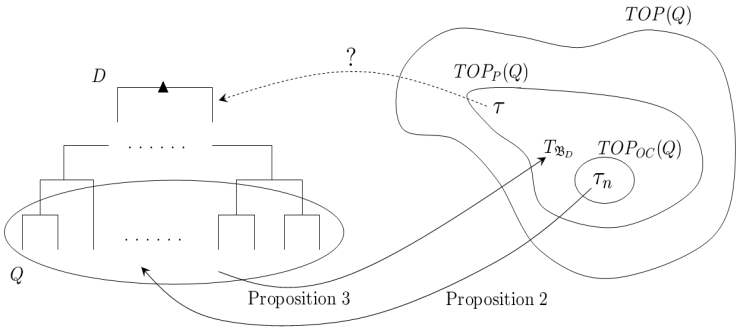

Figure 8: New family of topologies ($TOP_P(Q)$). The dashed arrow between $TOP_P(Q)$ and D rises the question on the possibility of obtaining a dendrogram from a topology $\tau \in TOP_P(Q)$.

Proposition 5. Let Q be a non-empty finite set and $\tau \in TOP_P(Q)$, then there exists

a dendrogram D defined over Q and a family $\mathfrak{B}_D$ of subtrees of D, such that $T_{\mathfrak{B}_D} = \tau$.

Proof. Note that it is sufficient and necessary to find a dendrogram D defined on Q such that for all $x \in Q$, O_x is a subtree of D. Hence, we can consider $\mathfrak{B}_D = \{O_x \in \tau \mid x \in Q\}$. In such a case $\mathfrak{B}_D$ covers Q and it is also a family of subtrees; in addition, it is a basis for τ. Then, the topology built up starting from this family $\mathfrak{B}_D$ using Proposition 3 is the same than the former topology, that is $T_{\mathfrak{B}_D} = \tau$.

We use induction over the cardinality of Q in this proof. If $|Q| = 1$ then the only possible topology on Q is $\mathcal{P}(Q)$ and this topology can be built starting from the only possible dendrogram on Q. Suppose the proposition is true for any set of cardinality lower or equal to n, that is: If $|Q| = k \leq n$ and τ is a topology in $TOP_P(Q)$ then there is a dendrogram D defined on Q and a family of subtrees $\mathfrak{B}_D$ covering Q such that $T_{\mathfrak{B}_D} = \tau$.

Now, we prove the proposition for a set Q with $n+1$ elements, it is $|Q| = n+1$. Let $\tau \in TOP_P(Q)$ and consider the family $\mathcal{B} = \{O_x \in \tau \mid x \in Q\}$, then $\mathcal{B} = \{B_1, B_2, \ldots, B_m\}$ where $B_1, B_2, \ldots, B_m$ are different subsets of Q. There are two possibilities, that one of them is Q or not.

Case 1. Without loss of generality we can consider that $B_m = Q$. Consider $A = B_1 \cup B_2 \cup \cdots \cup B_{m-1}$, then $A \neq Q$, A is an open set and suppose $A \neq \varnothing$. If $x \notin A$ then $O_x = Q$. Consider the topology $\tau\mid_A$ (restricted topology to A). We can see that $\tau\mid_A \in TOP_P(A)$ and $1 \leq |A| < |Q|$, it is $|A| \leq n$. Thus, there exist a dendrogram D_A defined over A such that $\tau\mid_A$ can be generated through subtrees of this dendrogram. Let build a dendrogram D_{A^C} on A^C and join these two dendrograms together by their root nodes to a new root node. Hence, we build a new dendrogram D over $A \cup A^C = Q$.

Let us use the initial observation to prove that there exists a subfamily $\mathfrak{B}_D$ of subtrees of D covering Q in such a way that $\tau = T_{\mathfrak{B}_D}$. Let $x \in Q$, then either $x \in A$ or $x \in A^C$; if $x \in A$ then owing to A is an open set, then $O_x \subseteq A$ and for this reason O_x is a subtree of the dendrogram D_A therefore a subtree of the dendrogram D. If $x \in A^C$ then $O_x = Q$, which corresponds to the whole dendrogram D. For this reason, $\mathcal{B}$ is a family $\mathfrak{B}_D$ of subtrees of D. In the case that $A = \varnothing$ then $\tau = \{\varnothing, Q\}$ which is constructed from any dendrogram. In this way the case 1 ends.

Case 2. If $B_i \neq Q$ for all $i = 1, 2, \ldots, m$, let $x \in Q$. It is impossible that $x \in B_1 \cap B_2 \cap \cdots \cap B_m$ since it implies that for every pair B_i, B_j then $B_i \cap B_j \neq \varnothing$ and for this reason $B_i \subseteq B_j$ or $B_j \subseteq B_i$. The consequence of this is that there is a B_i containing $B_1, B_2, ..., B_m$, for instance B_m. Thus, $B_1 \cup B_2 \cup \cdots \cup B_m = B_m$ but this family covers Q, it is $B_1 \cup B_2 \cup \cdots \cup B_m = Q$, thus $B_m = Q$ an this result contradicts our hypothesis. Let us consider A as the union of the B_i that contains a fix element $x_0 \in Q$, it is $A = \cup\{B_i \in \mathcal{B} \mid x_0 \in B_i\}$. We know A is an open set, $A \neq \varnothing$, $A \neq Q$ and $A^C = \cup\{B_i \in \mathcal{B} \mid x_0 \notin B_i\}$, then A^C is also an open set since it is an union of open sets. Consider again the topologies restricted to each set, $\tau\mid_A$ and $\tau\mid_{A^C}$. Once again $\tau\mid_A \in TOP_P(A)$ and $\tau\mid_{A^C} \in TOP_P(A^C)$, furthermore $1 \leq |A| < |Q|$ and $1 \leq |A^C| < |Q|$,

it is $|A| \leq n$ and $|A^C| \leq n$, then for the induction hypothesis there exist dendrograms D_A and D_{A^C} generating these topologies. Now let join these dendrograms together by their root nodes to build up a dendrogram D. Again, we will use the initial observation to conclude this proof; let $x \in Q$, then either $x \in A$ or $x \in A^C$, if $x \in A$ then $O_x \subseteq A$ owing to A is an open set, then O_x is a subtree of D_A and for this reason a subtree of D. If $x \in A^C$ the result is analogous. □

Proposition 5 shows that the family $TOP_P(Q)$ is the collection of "all" possible topologies that can be obtained from dendrograms. Thus, every topology not belonging to this family cannot be represented by means of a dendrogram; an example of this is the included point topology. Suppose that Q has more than two elements and also suppose that the topology τ_a, defined by $\tau_a = \{O \subseteq Q \mid a \in O\}$ where $a \in Q$, can be obtained by means of a dendrogram D. The root node of D shows the existence of two large subtrees, in one of those subtrees ought to be a. Consider an element $b \neq a$ such that b can be found in the other subtree in such a way that the lowest subtree containing a and b is the whole dendrogram D, it means Q (Figure 9). On the other hand, $O_b = \{a, b\}$ given this topology. Thus, $Q \subseteq O_b$ for Proposition 5, which is a contradiction because Q has more than two elements.

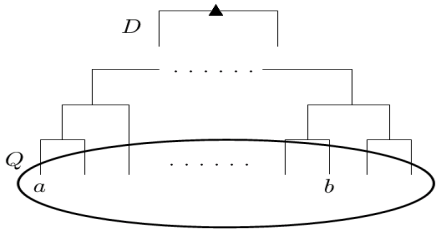

Figure 9: Two elements in Q that are in two disjoint subtrees covering Q.

Another interpretation of Proposition 5 is that for each topology in $TOP_P(Q)$ there exists a dendrogram generating such a topology, even more, when proving Proposition 5 we showed how that dendrogram may be built up. The question arises whether this dendrogram is unique. The answer is *no* and is also *no* for topologies obtained through maximal n-subtrees (Proposition 2). The reason is that a topology is determined by the open sets of a topological basis, which in turn have an associated dendrogram. This dendrogram might be whichever one mirroring the similarities expressed by the open set, and the number and type of such dendrograms increases rapidly with the cardinality of the open set [15, 17].

Finally, we have observed the following relations among the families of topologies mentioned in this paper: $TOP_{OC}(Q) \subset TOP_P(Q) \subset TOP(Q)$ (strictly contained). The last remark shows the relations between the two families characterised in this paper.

5 Conclusions

We presented a new methodology for constructing topologies over a non-empty finite set Q using dendrograms. In this way, the chemotopological methodology already existent was generalised. This usual chemotopological method is based on the construction of a topology for Q using maximal n-subtrees, which in turn depends on the selection of an integer n.

The generalisation developed in this paper makes more flexible the usual chemotopological method since now the construction of topologies is not restricted to the selection of an integer n. This generalisation keeps the sense of the chemotopology, that is the idea of building topologies using similarity information gathered in a dendrogram. Additionally, because of the consideration of all possible neighbourhoods (not restricted by cardinality) of the elements in Q, then all the similarity relationships take part in the construction of the topological basis, therefore this basis collects local similarities and also similarities among neighbourhoods. On the other hand, we have completely characterised the family of topologies that can be obtained from dendrograms and that also contains similarity information of the set Q under study. In this characterisation the main criterion is the use of subtrees to define the topological basis. The use of subtrees is justified since they represent similarity relationships among the elements under study. Finally, the theory developed in this paper can be considered, so far, the most general one in the context of building topologies from dendrograms.

6 Acknowledgements

The authors thank the Universidad del Valle and the Universidad de Pamplona for the financial support given during this research.

References

[1] Restrepo, G.; Mesa, H.; Llanos, E. J.; Villaveces, J. L. Topological study of the periodic system. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 68-75.

[2] Restrepo, G.; Llanos, E. J.; Mesa, H. On the topological sense of chemical sets. *J. Math. Chem.* **2006**, *39*, 363-376.

[3] Restrepo, G.; Mesa, H.; Llanos, E. J.; Villaveces, J. L. Topological study of the periodic system. In *The mathematics of the periodic table*; King, B.; Rouvray, D., Eds.; Nova: New York, 2006; Chapter 5, 75-100.

[4] Restrepo, G.; Llanos, E. J.; Mesa, H. Topological space of the chemical elements and its properties. *J. Math. Chem.* **2006**, *39*, 401-416.

[5] Restrepo, G.; Villaveces, J. L. From trees (dendrograms and consensus trees) to topology. *Croat. Chem. Acta*, **2005**, *78*, 275-281.

[6] Daza, M. C.; Restrepo, G.; Uribe, E. A.; Villaveces, J. L. Quantum chemical and chemotopological study of fourth row monohydrides. *Chem. Phys. Lett.* **2006**, *428*, 55-61.

[7] Cupal, J.; Kopp, S.; Stadler, P. F. RNA shape space topology. *Artif. Life* **2000**, *6*, 3-23.

[8] Stadler, B. M. R.; Stadler, P. F.; Wagner, G. P.; Fontana, W. The topology of the possible: Formal spaces underlying patterns of evolutionary change. *J. Theor. Biol.* **2001**, *213*, 241-274.

[9] Stadler, B. M. R.; Stadler, P. F.; Shpak, M.; Wagner, G. P. Recombination spaces, metrics, and pretopologies. *Z. Phys. Chem.* **2002**, *216*, 217-234.

[10] Stadler, B. M. R.; Stadler, P. F. Generalized topological spaces in evolutionary theory and combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 577-585.

[11] Stadler, B. M. R.; Stadler, P. F. The topology of evolutionary biology. In *Modelling in molecular biology*; Ciobanu, G.; Rozenberg, G., Eds.; Springer Verlag, Natural Computing Series: Berlin, 2004; 267-286.

[12] Gordon, A. D. *Classification*; Chapman and Hall: London, 1981.

[13] Everitt, B. S. *Cluster analysis*; Arnold: London, 1995.

[14] Potter, M. *Set theory and its philosophy*; Oxford: Oxford, 2004.

[15] Felsenstein, J. The number of evolutionary trees. *Syst. Zool.* **1978**, *27*, 27-33.

[16] Restrepo, G.; Mesa, H.; Llanos, E. J. Three dissimilarity measures to contrast dendrograms. *J. Chem. Inf. Model.* **2007**, *47*, 761-770.

[17] Wedderburn, J. H. M. The functional equation $g(x^2) = 2\alpha x + [g(x)]^2$. *Ann. Math.* **1922**, *24*, 121-140.