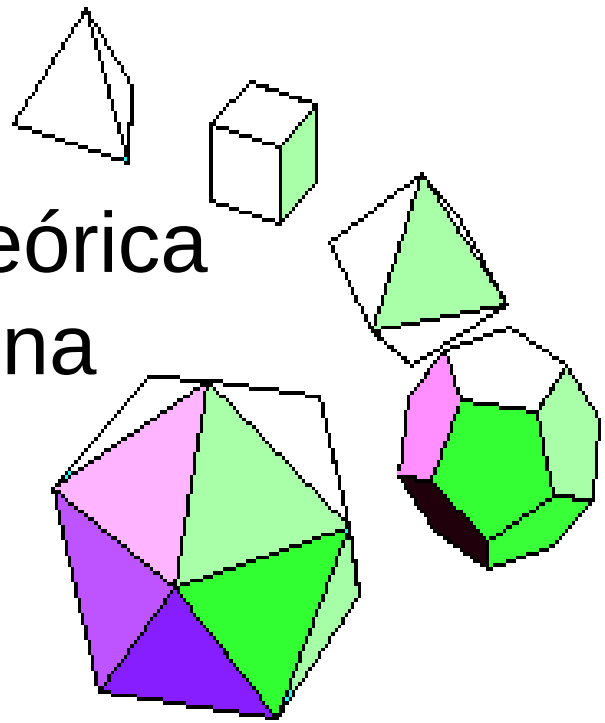
FROM DENDROGRAMS TO TOPOLOGY

Guillermo Restrepo
and
José Luis Villaveces

Laboratorio de Química Teórica
Universidad de Pamplona
Pamplona
Colombia



ABSTRACT

We describe a methodology to starting from dendrograms and consensus trees provide with a topology a set of chemical interest. We show four practical examples of the methodology (72 chemical elements, 31 steroids, 250 benzimidazoles and 20 amino acids) and besides the study of some topological properties such as closures and boundaries.

INTRODUCTION

There are several chemical systems which are characterised for the similarity relationships among their elements (chemical elements, acids, bases and so on). One way to quantify this similarity starts with the representation of every chemical object as a vector of its properties(1). Then, it is calculated the similarity among all vectors by means of a similarity function (1).

INTRODUCTION

A methodology that has shown important results trying to find similarities is the cluster analysis, which, taking advantage of several grouping methodologies, finally shows clusters of elements that share common features (2). A way to visualise such clusters is a 2D graphic representation called *dendrogram* (a tree in mathematical terms (3)) whose branches show groups of similar elements. Cluster analysis finish obtaining a dendrogram and interpreting it. But as we show (4,5), it is possible to interpret a dendrogram as a map of neighbourhoods of the elements; and extracting of these clusters the notion of neighbourhood.

INTRODUCTION

We can approach this interpretation and apply the mathematical theory in charge of study neighbourhood relationships (*topology*) (6). With this tool it is possible to define topologies on the set and to study some topological properties of itself, as closures and boundaries among others. In this work we show 4 practical examples of our methodology (bezimidazoles, amino acids, steroids and chemical elements).

METHODOLOGY

Cluster Analysis

This methodology considers every element as a point (vector) (1) in a space of properties and calculate similarity among elements according to several mathematical functions (metrics and non-metrics) (7). After, it grouping elements taking advantage of the concept of distance between an element and a set (Figure 1). Finally, clusters of elements are shown in a dendrogram (Figure 2a).

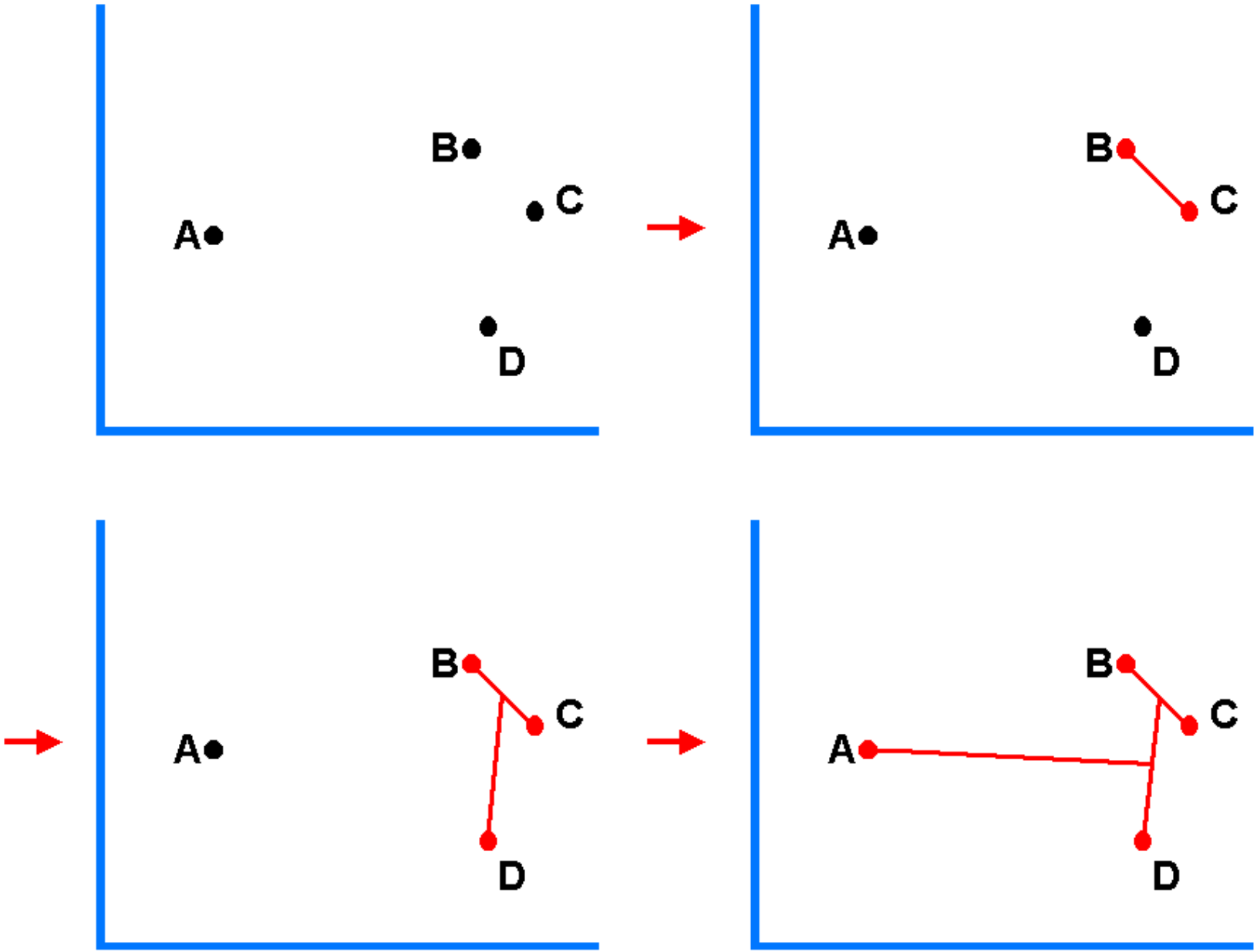


Figure 1

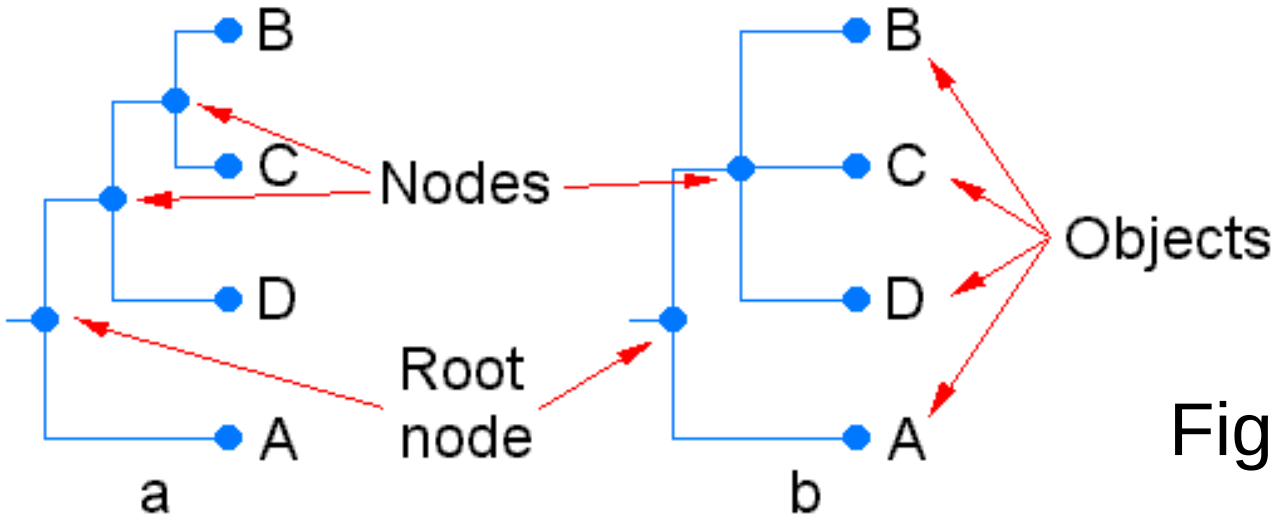


Figure 2

However, a selection of a particular similarity function and grouping methodology introduce an arbitrariness in the study, for this reason is recommendable to do consensus trees (4,5) (Figure 2b). Thus, dendrograms and consensus trees can be defined, in general, as trees:

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 (Figure 2).

With the aim of providing a topology on the set of chemical interest we introduce the following definitions:

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

1. *G* does not contain the root node;

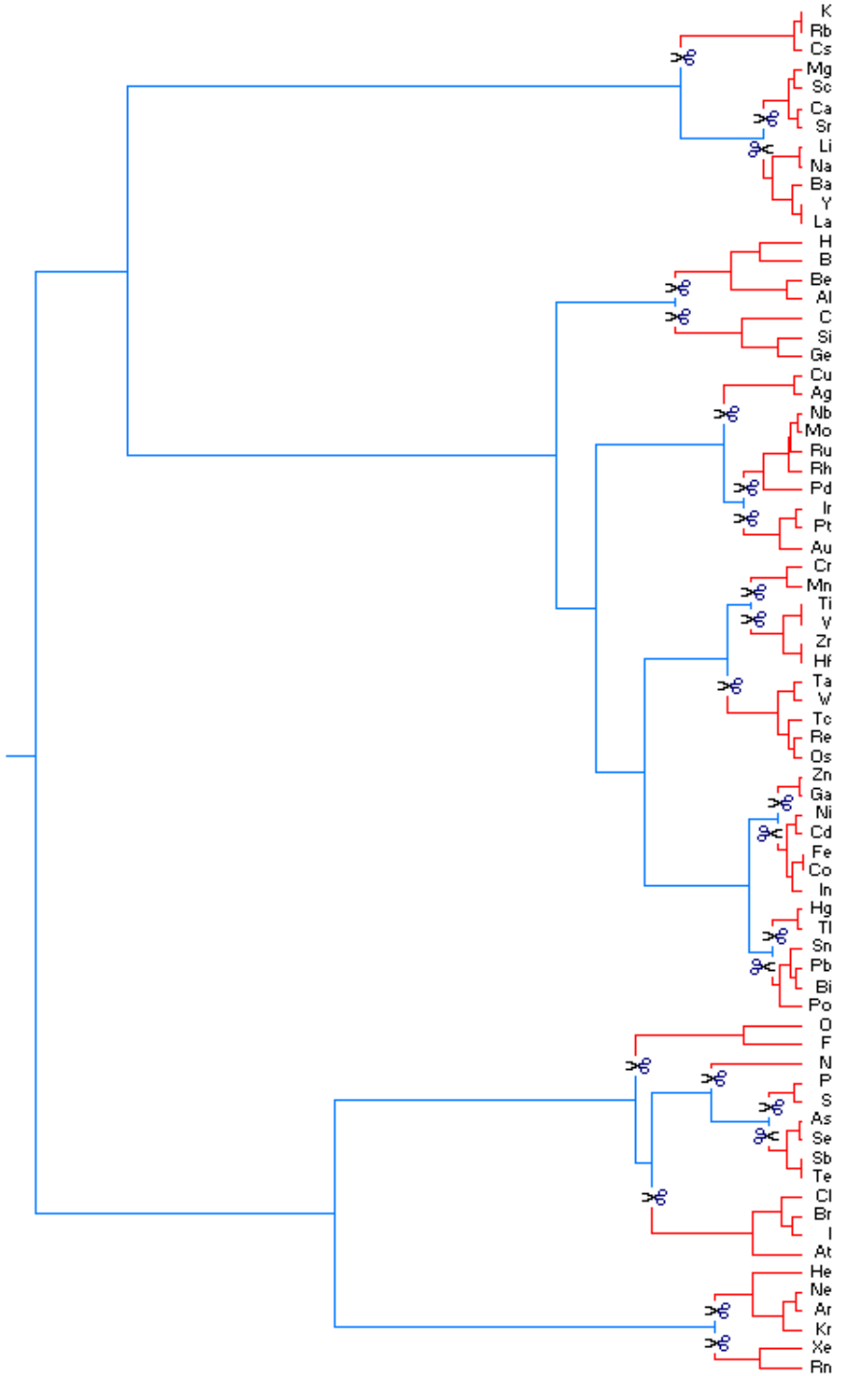
2. There is a node *p* of *T* with degree greater than 1 such that *G* corresponds to one of the connected subgraphs obtained subtracting *p* from *D*.

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.

Figure 3:

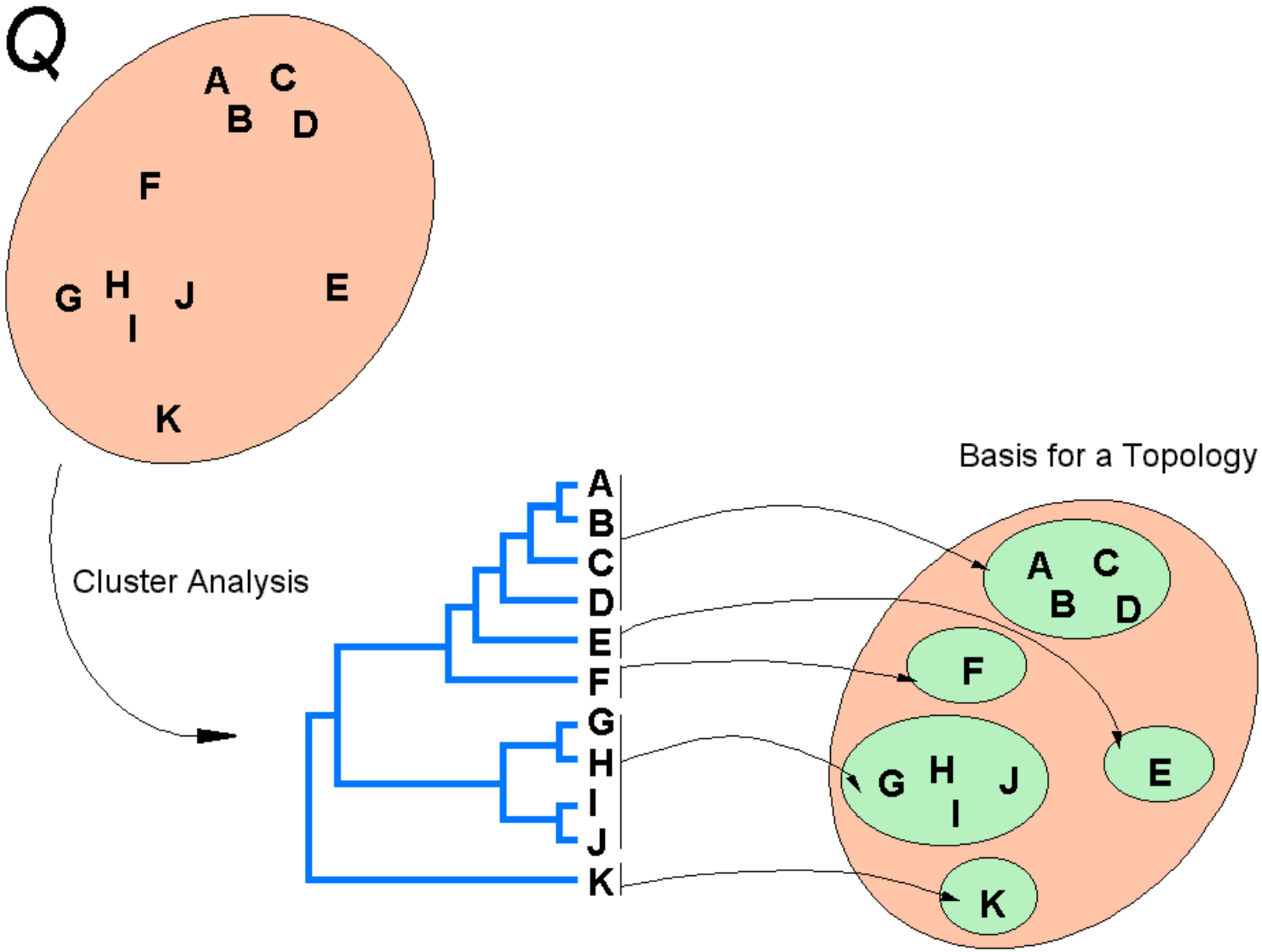
Maximal 5-subtrees



We build up a basis for a topology using the following theorem:

Theorem 1: Let Q be a set of chemical interest and $\boldsymbol{B_n} = \{B \subseteq Q \mid B$ be formed by elements of some maximal n-subtree$\}$. Then, $\boldsymbol{B_n}$ is basis for a topology $\blacklozenge_n$ on Q (A3).

Thus n is 1⊙n ⊙$\#_Q$


Q
A
C
B
D
F
G
H
J
E
I
K
Cluster Analysis
A
B
C
D
E
F
G
H
I
J
K
Basis for a Topology
A
C
B
D
F
E
G
H
J
I
K

Once we have provided the set Q with a topology $\blacklozenge_n$ we can study some topological properties on the set Q.

Let $A \subset Q$ and $x \in Q$, it is said that x is a **closure point** of A iff for every $O \in \tau_n$, such that $x \in O$, then $O \cap A \neq \varnothing$

Let $A \subset Q$, the **closure** of A is defined as:

$$A = \{ x \in Q \mid x \text{ is closure point of } A \}$$

Let $A \subset Q$ and $x \in Q$, it is said that x is an **accumulation point** of A iff for every $O \in \tau_n$, such that $x \in O$, then

$$(O - \{x\}) \cap A \neq \varnothing$$

Let $A \subset Q$, the **derived set** of A is defined as:

$$A' = \{ x \in Q \mid x \text{ is accumulation point of } A \}$$

Let $A \subset \mathbb{Q}$ and $x \in \mathbb{Q}$, it is said that x is a **boundary point** of A iff for every $O \in \tau_n$, such that $x \in O$, then $O \cap A \neq \emptyset$ and $O \cap (\mathbb{Q} - A) \neq \emptyset$

Let $A \subset \mathbb{Q}$, the **boundary** of A is defined as:

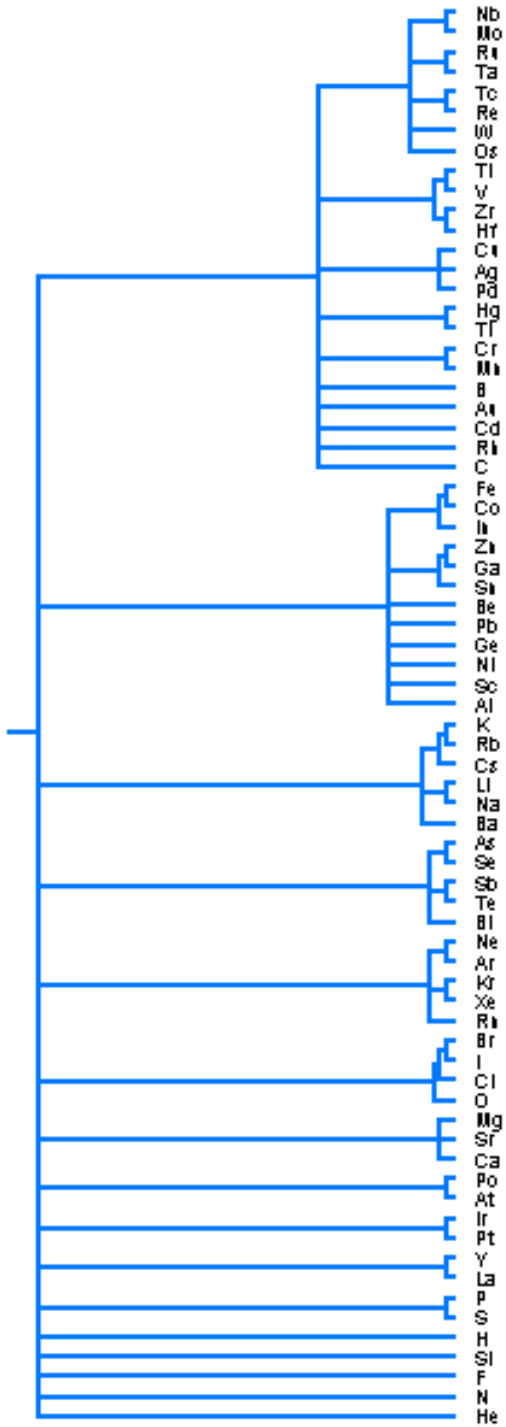
$$b(A) = \{x \in \mathbb{Q} \mid x \text{ is boundary point of } A\}$$

PRACTICAL EXAMPLES OF CHEMICAL INTEREST

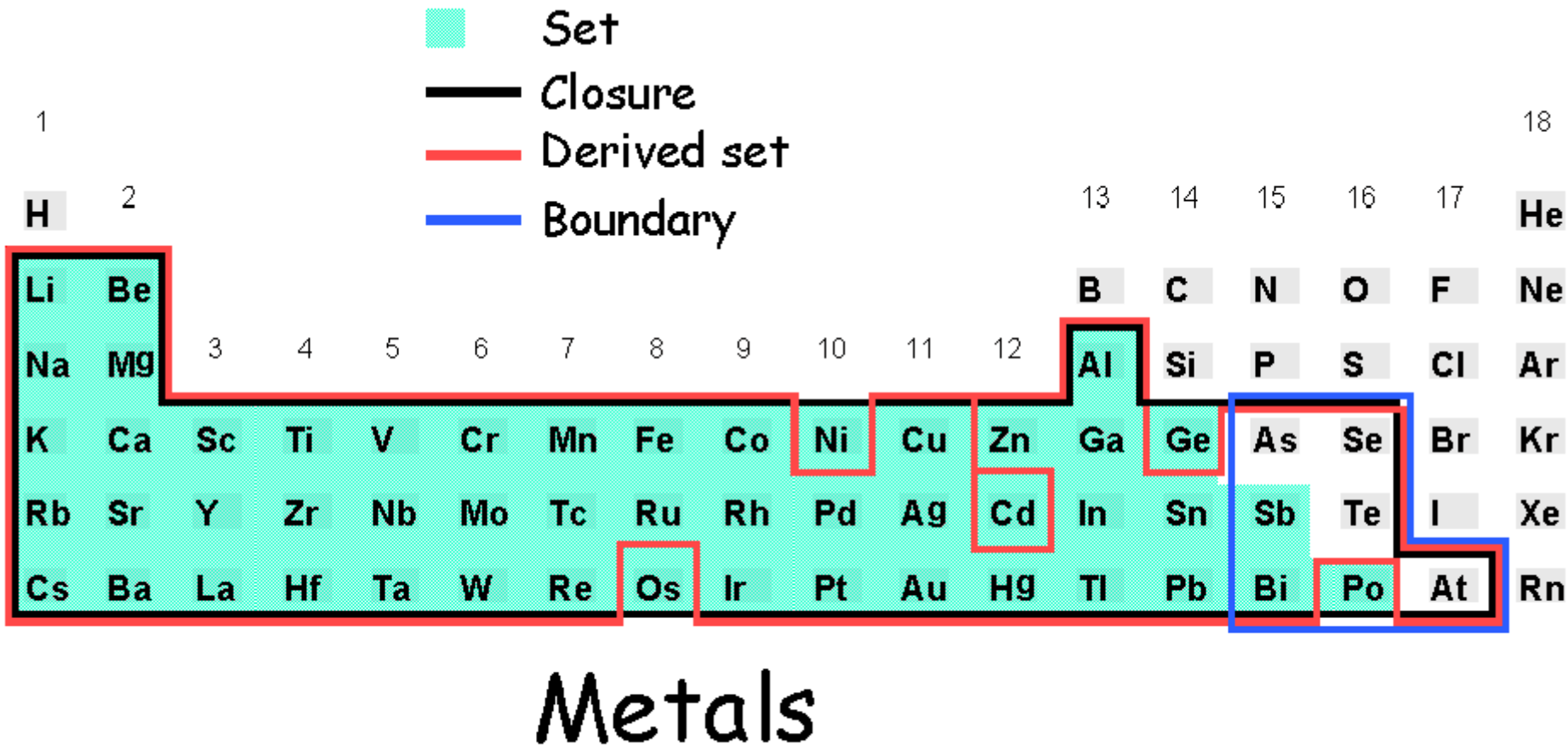
CHEMICAL ELEMENTS

We build up (4,5) a topology $\blacklozenge_5$ on the set *Q* of 72 chemical elements (Z=1-86, omitting 58-71) every one defined by 31 physico-chemical properties.

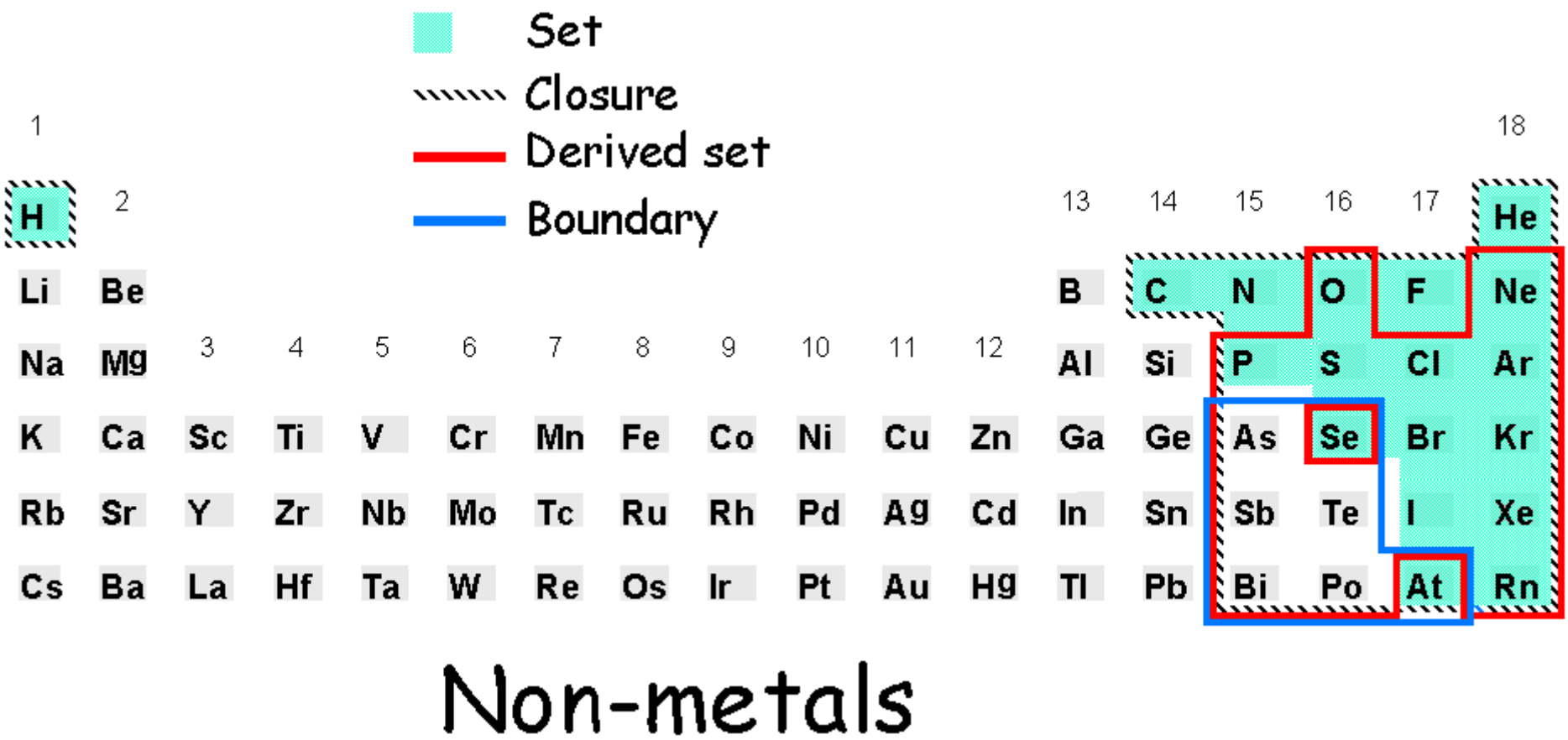
Adams consensus



$$B_5 = \left\{ \begin{array}{l} \{Ti,V,Zr,Hf\},\{Ru,Ta\},\{Tc,Re\},\{Nb,Mo\},\{W\},\{Os\},\{Cu,Ag,Pd\}, \\ \{Hg,Tl\},\{Cr,Mn\},\{B\},\{Au\},\{Cd\},\{Rh\},\{C\},\{Fe,Co,In\},\{Zn,Ga,Sn\} \\ \{Be\},\{Pb\},\{Ge\},\{Ni\},\{Sc\},\{Al\},\{K,Rb,Cs\},\{Li,Na\},\{Ba\}, \\ \{As,Se,Sb,Te,Bi\},\{Ne,Ar,Kr,Xe,Rn\},\{Br,I,Cl,O\},\{Mg,Sr,Ca\}, \\ \{Po,At\},\{Ir,Pt\},\{Y,La\},\{P,S\},\{H\},\{Si\},\{F\},\{N\},\{He\} \end{array} \right\}$$


Set
Closure
Derived set
Boundary
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
H
He
Li
Be
B
C
N
O
F
Ne
Na
Mg
Al
Si
P
S
Cl
Ar
K
Ca
Sc
Ti
V
Cr
Mn
Fe
Co
Ni
Cu
Zn
Ga
Ge
As
Se
Br
Kr
Rb
Sr
Y
Zr
Nb
Mo
Tc
Ru
Rh
Pd
Ag
Cd
In
Sn
Sb
Te
I
Xe
Cs
Ba
La
Hf
Ta
W
Re
Os
Ir
Pt
Au
Hg
Tl
Pb
Bi
Po
At
Rn

Metals


Set
Closure
Derived set
Boundary
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
H
He
Li
Be
B
C
N
O
F
Ne
Na
Mg
Al
Si
P
S
Cl
Ar
K
Ca
Sc
Ti
V
Cr
Mn
Fe
Co
Ni
Cu
Zn
Ga
Ge
As
Se
Br
Kr
Rb
Sr
Y
Zr
Nb
Mo
Tc
Ru
Rh
Pd
Ag
Cd
In
Sn
Sb
Te
I
Xe
Cs
Ba
La
Hf
Ta
W
Re
Os
Ir
Pt
Au
Hg
Tl
Pb
Bi
Po
At
Rn

Non-metals

BENZIMIDAZOLES

Niño, Daza and Tello (8) developed a dendrogram (euclidean-single linkage) of 238 benzimidazoles using graph theoretical and quantum mechanical descriptors. In this set it is possible classify compounds according to their pharmacological activity: Angiotensin II (A), Antivirals (AV), Cardiotonics (C) and Antihelmintics (H). Cardinalities of every class are: #A=158, #AV=15, #C=32, #H=33.

We build up the basis $\boldsymbol{B_{15}}$ on the set.

$\overline{A} = A \leftarrow \{h_{31}\}$

$b(A)=\{a_{8a},a_{6a},a_{18},a_{12},a_{16},a_{13},a_{15},a_5,a_{11},a_4,a_3,a_{14},a_{10},h_{31},a_0\}$

$\overline{AV} = AV \leftarrow \{h_{42}\}$

$b(AV)=\{h_{42},av_{12},av_{10},av_{14},av_{15},av_{11},av_9,av_8,av_7,av_5,av_4,av_6,av_3,av_2,av_1\}$

$\overline{C} = C$

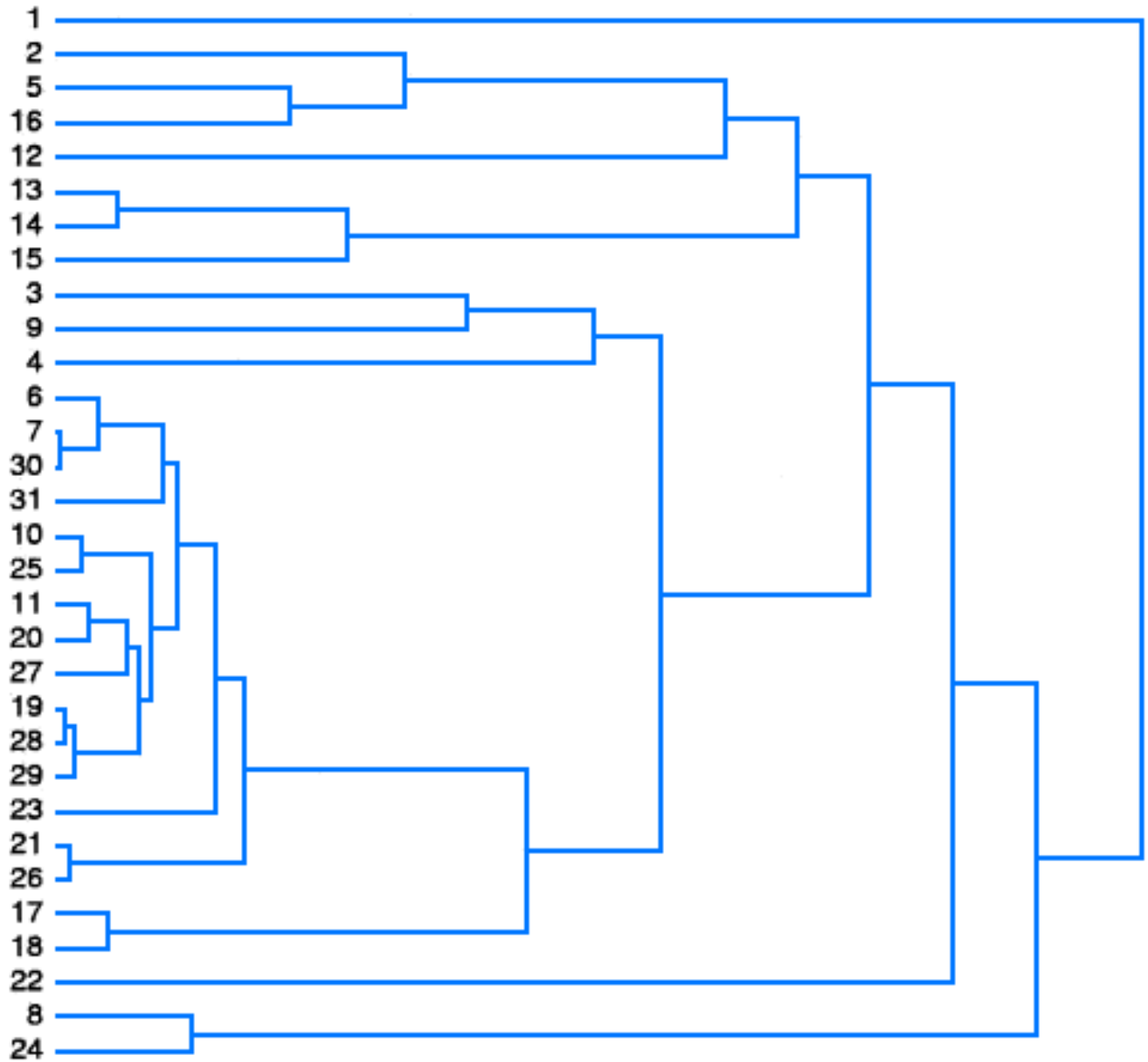
$b(C)=$ 

$$\bar{H} = H \Leftarrow \left\{ \begin{array}{l} av_{12},av_{10},av_{14},av_{15},av_{11},av_{9},av_{8},av_{7},av_{5},av_{4},av_{6},av_{3},av_{2}, \\ av_{1},a_{8a},a_{6a},a_{18},a_{12},a_{16},a_{13},a_{15},a_{5},a_{11},a_{4},a_{3},a_{14},a_{10},a_{0} \end{array} \right\}$$

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

STEROIDS

Bultinck and Carbó (9) developed a dendrogram (euclidean-stochastic transform) of 31 steroids using quantum descriptors. It is possible classify compounds in 5 groups according to chemical knowledge on structure and reactivity: 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).


1
2
5
16
12
13
14
15
3
9
4
6
7
30
31
10
25
11
20
27
19
28
29
23
21
26
17
18
22
8
24

$$B_2 = \left\{ \begin{matrix} \{1\},\{2\},\{3\},\{4\},\{5,16\},\{6\},\{7,30\},\{8,24\},\{9\},\{10,25\},\{11,20\},\{12\}, \\ \{13,14\},\{15\},\{17,18\},\{19,28\},\{21,26\},\{22\},\{23\},\{27\},\{29\},\{31\} \end{matrix} \right\}$$

Where 1=Aldosterone, 2=Androstanediol, 3=5-androstanediol, 4=4-androstanediol, 5=Androsterone, 6=Corticosterone, 7=Cortisol, 8=Cortisone, 9=Dehydroepiandrosterone, 10=11-deoxycorticosterone, 11=11-deoxycortisol, 12=Dihydrotestosterone, 13=Estradiol, 14=Estriol, 15=Estrone, 16=Ethiochonalonone, 17=Pregnenolone, 18=17a-hydroxypregnenolone, 19=Progesterone, 20= 17a-hydroxypregnenolone, 21=Testosterone, 22= Prednisolone, 23=Cortisolacetat, 24=4-pregnene-3,11,20-trione, 25=Epicorticosterone, 26=19-nortestosterone, 27=16a,17a-dihydroxyprogesterone, 28=17a-methylprogesterone, 29=Norprogesterone, 30=2a-methylcortisol, 31=2a-methyl-9a-fluorocortisol

$\bar{E} = E$

b(E)=

$\bar{W} = W$

b(W)=

$\bar{A} = A$

b(A)=

$\bar{D} = D$

b(D)=

$\bar{C} = C$

b(C)=

AMINO ACIDS

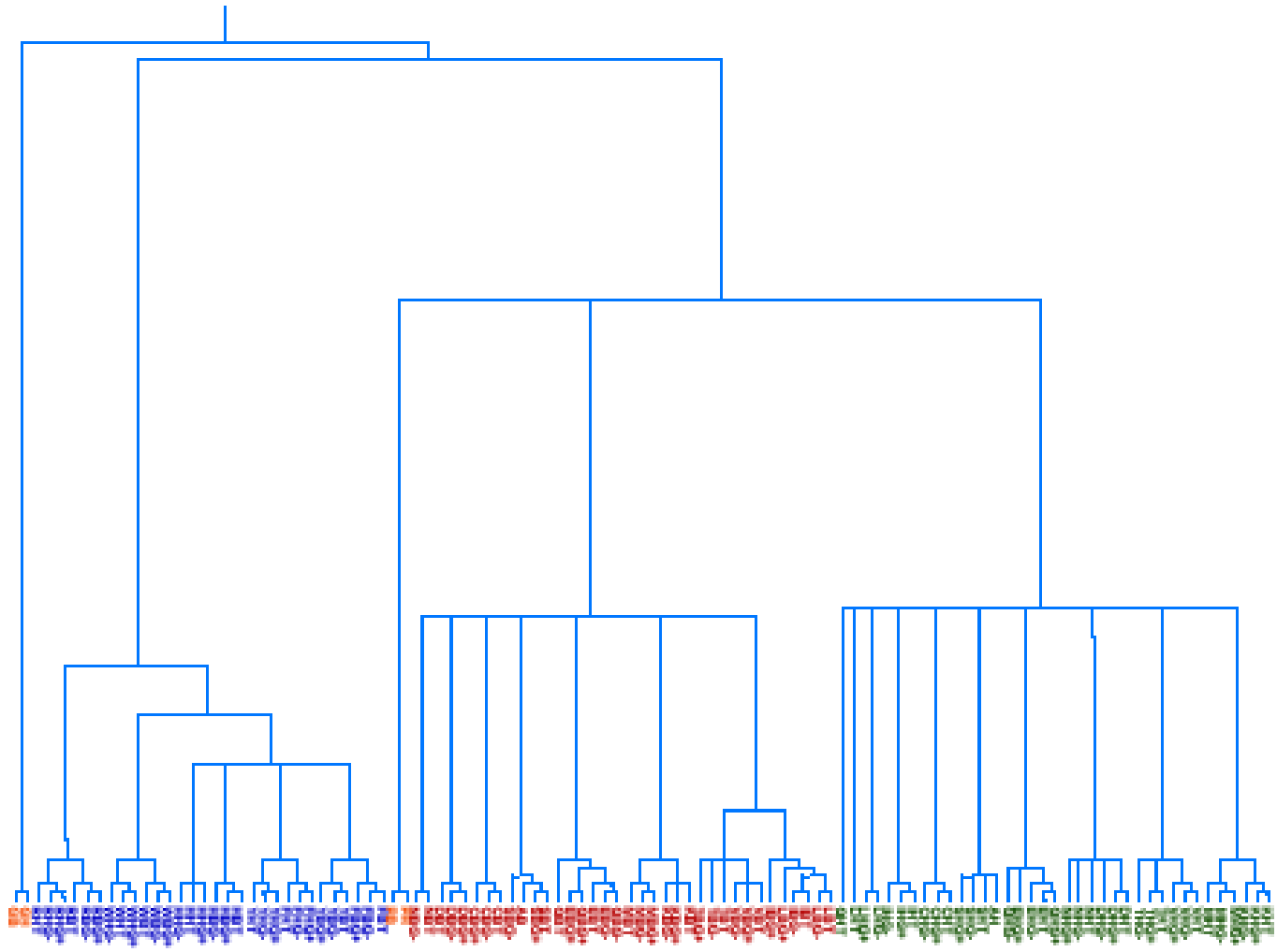
Authors (10) developed two dendrograms of 20 amino acids and 5 of their conformers using graph theoretical and quantum mechanical descriptors, after built up a consensus tree.

It is possible to classify amino acids as: amino acids with hydrophobic side groups (PHO), with hydrophilic side groups (PHI) and those that are in between (PP).

PHO={V,L,I,M,F}

PHI={N,E,Q,H,K,R,D}

PP={G,A,S,T,Y,W,C,P}



$$
B_7 = \left\{ \begin{array}{l}
\{kht, khg_+, khg_-, mhg_+, mhg_-, mht\}, \\
\{vht, vhg_-, vhg_+, tht, thg_+, thg_-\}, \\
\{dht, dhg_+, dhg_-, nht, nhg_-, nhg_+\}, \\
\{qhg_+, eht, ehg_-, ehg_+, qhg_-\}, \{ah, sht, shg_-, shg_+\}, \\
\{cht, chg_+, chg_-\}, \{iht, ihg_+, ihg_-\}, \{lhg_-, lhg_+\}, \{lht\}, \{qht\}, \\
\{lbt, lbg_-, ibt, ibg_-, lbg_+, ibg_+\}, \{vbt, vbg_+, vbg_-, tbt, tbg_-, tbg_+\}, \\
\{kbt, kbg_+, kbg_-, mbg_+, mbg_-, mbt\}, \\
\{qbg_-, ebg_-, ebt, qbg_+, ebg_+, qbt\}, \\
\{sbt, sbg_+, ab, sbg_-\}, \{cbt, cbg_-, cbg_+\}, \{dbg_+, dbg_-, dbt\}, \\
\{nbt, nbg_-\}, \{ph, pb\}, \{fbt, fbg_-, fbg_+, ybt, ybg_-, ybg_+\}, \\
\{fhg_+, fhg_-, fht, yhg_+, yhg_-, yht\}, \{hbt, hbg_+, hbg_-\}, \\
\{hhg_+, hht, hhg_-\}, \{wbg_+, wbg_-, wbt, whg_+, wht, whg_-\}, \\
\{rbg_+, rbg_-, rbt, rhg_+, rhg_-, rht\}, \{gh, gb\}
\end{array} \right\}
$$

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

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

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

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

$$\overline{PHI} = PHI \cup \left\{ \begin{array}{l} mgh_+, mgh_-, mht, mbg_+, mbg_-, mbt \end{array} \right\},$$

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

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

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

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

CONCLUSIONS

This methodology 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, very related to the chemical knowledge.

Chemical elements

We found relationships among elements already known (groups, diagonal relationships, inert pair effect) and some not known.

The mathematical boundary of the set of metals and non-metals is made of semimetals.

Benzimidazoles:

Classification of benzimidazoles in 4 classes does not give 4 disjuncts subsets, due to there are some benzimidazoles h and av which appear on several sets. But, it is correct to say that the set of cardiotonics is a group well defined because its closure is itself and its boundary is empty; this result shows that this set is disjunct in the space of benzimidazoles. On the other hand, according to our results is possible to speculate and say that those substances which are in the boundary of two subsets can have properties intermediate between two subsets.

Steroids:

All five subsets result be themselves their own closure ergo their boundaries were empty. These results indicate that this classification of steroids according to intuitive chemical knowledge on structure and reactivity gives disjunct sets, or in other words, robust groups.

Amino acids:

The closure of 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, in the closure appears all threonine and tyrosine amino acids, which are amino acids that are in between hydrophobic and hydrophilic ones.

The closure of subset of amino acids with hydrophilic side groups is build up, besides 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 two above is build up, besides 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.

REFERENCES

(1) Willet, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38,* 983-996.

(2) Barnard, J. M. and Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J . Chem. Inf. Comput. Sci*. **1992**, *32*, 644-649.

(3) Hendy, M. D. and Penny, D. Cladograms Should Be Called Trees. *Syst. Zool.* **1984**, *33*, 245-247.

(4) 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.

(5) Restrepo, G.; Mesa, H.; Llanos, E. J.; Villaveces, J. L. Topological Study of the Periodic System. In: King, B.; Rouvray, D. The Mathematics of the Periodic Table, in press, Nova Publishers, New York.

(6) Mendelson, B. *Introduction to Topology*. 3rd edition. New York: Dover, 1990. p. 1-28.

(7) Otto, M. *Chemometrics: Statistics and Computer Application in Analytical Chemistry*; Ed.; Wiley-VCH: Weinheim, 1999; pp 148-156.

(8) Niño, M.; Daza, E. E. and Tello, M. A Criteria To Classify Biological Activity of Benzimidazoles from a Model of Structural Similarity. *J. Chem. Inf. Comput. Sci.* **2001**, *41,* 495-504.

(9) Bultinck, P. and Carbó-Dorca, R. Molecular Quantum Similarity Matrix Based Clustering of Molecules Using Dendrograms. *J. Chem. Inf. Comput. Sci.* **2003**, *43,* 170-177.

(10) Villaveces, J. L.; Cárdenas, C.; Obregón, M.; Llanos, E. J.; Bohórquez, H.; Machado, E.; Patarroyo, M. E. Constructing a useful tool for characterizing amino acid conformers by means of quantum chemical and graph theory indices. *Comput. Chem. C* **2002**, *26*, 631-646.