**MATCH**
*Communications in Mathematical and in Computer Chemistry*

MATCH Commun. Math. Comput. Chem. **72** (2014) 453-474

ISSN 0340 - 6253

# Pretopological Spaces as a Classification Tool for RNAs Represented as a Succession

**Johan F. Galindo[†,§], Gustavo Rubiano[‡], Edgar E. Daza[§]**

[†] *Quantum Theory Project, Department of Chemistry, University of Florida, Gainesville, Florida, 32611, USA*

[‡] *Departamento de Matemáticas, Universidad Nacional de Colombia, Bogotá, Colombia*

[§]*Grupo de Química Teórica, Departamento de Química, Universidad Nacional de Colombia, Bogotá, Colombia*

(Received February 3, 2014)

### Abstract

RNA representation is one of the most important issues in bioinformatics. A challenge, however, is achieving a correct representation that is able to encode the directionality, secondary and primary structure. One possible solution is the RNA graph representation and its associate tree; however, it fails to encode the directionality of the RNA molecule. Therefore, a succession based on the tree representation is proposed as an alternative. From this novel succession, we have defined three metrics in order to compare the similarity between RNAs. Finally, we have used the properties of pretopological spaces which were generated from the metrics as a cluster tool. This methodology has allowed the classification of an RNA hairpin, which was mutated in two different positions.

## 1 Introduction

Ribonucleic Acids (RNAs) are among the most important molecules in life chemistry [1, 2]. RNA is of great importance for protein synthesis since it functions as a regulatory molecule and fulfills the same role as deoxyribonucleic acid (DNA) in some microorganisms, i.e., encoding genetic information [3–5].

One of the main challenges when working with systems involving biomolecules is to find an accurate representation that allows the comparison of two or more biomolecules [6–8]. The chosen representation must encode the information and the properties taken into account within the concept of chemical structure [9, 10]. Comparisons are limited by the degree of indeterminacy introduced by the unique features defined to establish the differences and similarities among the molecules of interest [11–13].

A previous model for both representing and comparing tRNAs has been developed by our group. [14] In this work, Sections 2 and 3 are an overview of the most common RNA representations and similarity measurements. A new RNA representation model is proposed taking into account the directionality and secondary structure of the molecule, as well as the electrostatic environment for every base pair derived from quantum electronic structure calculations (Section 4). These quantum mechanical calculations indirectly allow to incorporate some local elements of the tertiary structure, since the calculation that represent a particular nucleotide depends on the molecular environment defined by its nucleotide neighbors in typical frames. For this new representation, metric measurements were defined and proved, allowing the comparison of the molecules represented as mathematical objects (Section 5). However, the metric measures only allow comparisons between pairs of elements, and a comparison between more elements requires the usual algorithms of clustering; therefore, the use of pretopological spaces associated with the metrics is appropriated to establish a classification method, which yielded rigorous criteria for setting up classification compared to the usual clustering techniques (Section 6). Finally, an application assay is done in order to show the power of this novel methodology (Section 7).

## 2 RNA Graph Representations

The graph theory has been used to address the challenge of modeling RNA structures and comparing them. Mathematically, a graph is seen as an object composed of a collection of points (vertices or simplex) $\{V_i\}$ and a set of unordered couples (links or edges) $E(G) = \{(V_i, V_j)\} = \{e_{ij}\}$, corresponding to the relationship between points $V_i$ and $V_j$ [15–18].

There are different graph representations depending on the type of vertices and edges assigned, ranging from a very simplified representation where the number of vertices is minimal to those where each nucleotide is a vertex (Figure 1.).

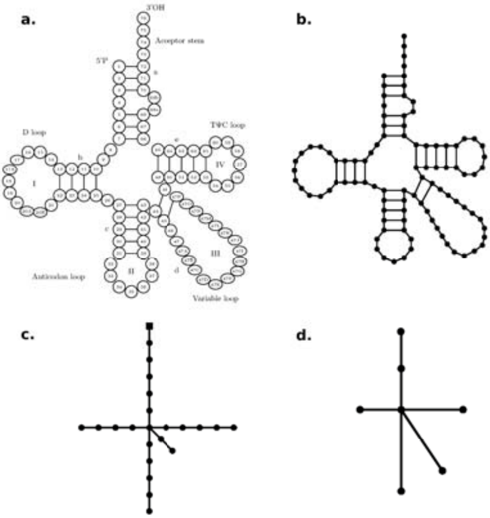

Figure 1: **a.** RNA secondary structure. **b.** RNA graph representation resulting from construction 1. Each nucleotide is a vertex and each hydrogen bond or phosphodiester bond corresponds to an edge. **c.** Contact structure associated with the contact matrix for the secondary RNA structure, obtained using construction 2. **d.** Secondary graph representation, where each edge corresponds to a double helix region and each vertex corresponds to a loop.

*Construction 1.*

For any RNA sequence with $n$ nucleotides, the set of vertices $V_i$ is formed by each of the nucleotides in the sequence and the set of edges is made up by the different types of bonds between nucleotides (Figure 1b) [20].

The connectivity matrix is then formed as follows:

Let $\mathbf{A}$ be the matrix connectivity of graph $G = \{V_i, E\}$, where $E = \{(V_i, V_j)\}$ is the set of couples that:

$$a_{i,j} = a_{j,i} = 1 \quad \text{if } (V_i, V_j) \in E \;\forall\; 1 \leqslant i, j \leqslant n,$$

and the result is zero for all other cases. This representation is very accurate as it preserves most elements forming RNA as well as its secondary structure features [19].

*Construction 2.*

Another graph representation of the RNA structure can be obtained from the matrix of contacts where the RNA-associated adjacency matrix only takes into account hydrogen bond interactions and does not include the backbone region.

Let **B** be the matrix formed by the following elements:

$$a_{i,i+1} = a_{i+1,i} = 1 \quad \text{for } 1 \leqslant i \leqslant n,$$

hence, the matrix is formed by the backbone and all other elements will be equal to zero. Therefore, the contact matrix **C** will then be:

$$\mathbf{C} = \mathbf{A} - \mathbf{B}.$$

Consequently, the graph associated with this matrix will contain the vertices corresponding to the hydrogen bonds and the edges will be related to the double helix areas that allow bonding between the vertices [21, 22], (Figure 1c).

*Construction 3.*

This construction assigns a set of vertices $\{V_i\}$ as the set of loops within the RNA secondary structure. A loop is the region in the secondary structure similar in shape to a circle and consists of five or more nucleotides. The edges of the graph correspond to the double helix regions that allow bonding between loops (Figure 1d). This representation is the most simplified structure presented here [23–26].

*Construction 4.*

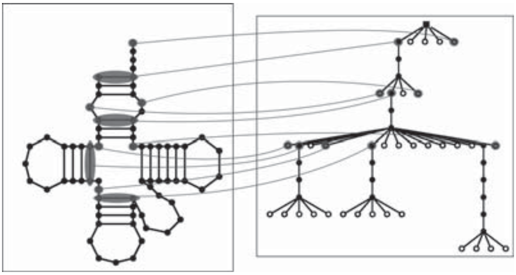

Figure 2: Tree structure obtained using construction 4. The black vertices correspond to a pair of nucleotides and the white ones correspond to unpaired nucleotides. The root vertex is virtual.

The tree representation was proposed by Schuster et al. [27]. In this representation, a pair of nucleotides corresponds to a black vertex, an unpaired nucleotide corresponds to a white one, and the root vertex (black square) is virtual (Figure 2). Constructions 1 to 3 have been widely used to study different aspects of RNA structure, [19] but for our proposal this last construction has been employed.

## 3 Comparing Graphs

Having established a RNA representation, the next step is to establish ways of comparing them. For such purpose, people have developed similarity measurements which are often metric measurements taking into account that the smaller the distance between molecules, the higher their degree of similarity.

Each distance depends on the type of representation used. One of the most frequently used representations comes from bioinformatics and is based on molecular similarity. Taking the RNA sequence into account, an editing metric reflects the number of operations necessary to go from one sequence to another. These operations are sometimes weighted by giving a higher value to one of them. These operations are known as nucleotide insertion, nucleotide deletion and nucleotide substitution.

Other metrics have been developed over other types of secondary structure representations. One of the most popular corresponds to *the base pair metric*. It is based on the bracket representation and was introduced by Zuker [28]. This metric has been widely used for secondary structure prediction. For tree-graph models a corresponding tree metric has been proposed [29], and for the mountain representation a mountain metric was developed [30]. Both metrics have also been used as a tool to predict the RNA secondary structures, but being less successful. Some other important metrics can be found in the work of Reidys and Stadler [21], where three different metric measurements are based on the structure of contacts, summarized as follows [22]:

- The $d_{srg}$ metric given by:

$$d_{sgr}(\Gamma_1,\Gamma_2) = (ln2)|Q_1 \Delta Q_2| \ .$$

- The $d_{inv}$ metric is equal to $d_{mag}$, defined as:

$$d_{inv}(\Gamma_1,\Gamma_2) = d_{mag}(\Gamma_1,\Gamma_2) = \frac{1}{ln2} d_{sgr}(\Gamma_1,\Gamma_2) - 2\Omega(\Gamma_1,\Gamma_2) \ .$$

Where $\Gamma_i$ corresponds to the contact structure associated with RNA, $Q_i$ corresponds to the corners of the structures, $|Q_1 \Delta Q_2|$ is the cardinality of the symmetrical difference $|Q_1 \cup Q_2| - |Q_2 \cap Q_1|$ and $\Omega(\Gamma_1, \Gamma_2)$ are the number of cycles in the graph $\Gamma_1 \Delta \Gamma_2$. It is worth highlighting that the compared RNAs must have the same number of nucleotides.

It is also possible to find some other metrics [21,22,31,32] that can compare different RNAs, but all of these metrics might be limited. One of their limitations is due to the way that they have been constructed, since some of these do not differentiate between a pair of *A-U* and *G-C* and thus simplify the graph to a collection of points.

# 4 From a Graph to a Succession

We started by considering the primary structure to keep RNA directionality in an explicit and useful way. Since RNA is a succession of nucleotides, we expected that the RNA's secondary structure could also be represented as a succession of terms. In other words, we are proposing to manage a Bi-dimensional problem, even a tertiary one, as a one dimensional problem. It can be achieved without loosing the features associated with its secondary structure and including quantum mechanical factor through the atomic partial charge distribution for each nucleotide.

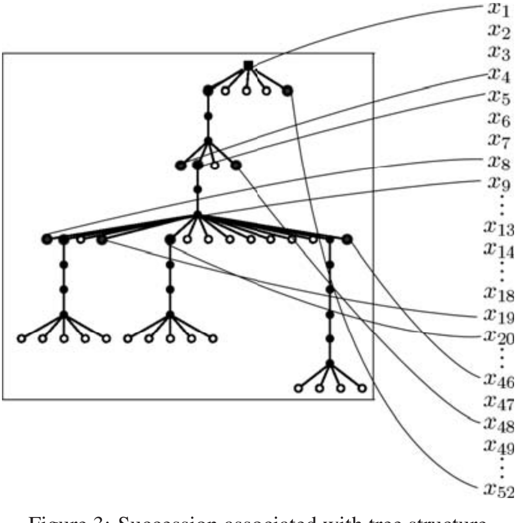

Figure 3: Succession associated with tree structure

The tree representations that correspond to the fourth construction have a well-established order, it is possible to further simplify them, giving rise to a succession of vertices as shown in

Figure 3. The construction of this representation begins by assigning the leftmost vertex below the first succession element. If the node is white, the next term of the succession corresponds to the node on its right and so on until finding a black node. Once a black node is found, the next term of the succession would correspond to the couple that is linked to it in the immediately lower level. As explained in the previous case, if the node is white, the terms are assigned by moving through the succession from left to right until reaching the next black node. The procedure is repeated until finishing with each branch. The next term in the succession would correspond to the node on the right side of the branch root.

In a previous representation developed by our group [14], each vertex was correlated to a convenient linear combination of partial charges associated with the most affected atoms in each nitrogenated base when its first neighbors are changed. This set of weights represents fundamental information in the succession. Briefly, in the case of a pair of nucleotides, the first four components corresponds to the combinations of partial atomic charges from the first four main components associated with one of the nucleotides and the last four correspond to the other nucleotide. This information represents the electrostatic environment of a particular base pair and it also brings the possibility to include a chemical description in the representation.

To take this type of information into account over the new representation, each element in the succession was assigned to an 8-tuple. For each black point in the tree representation (pair of nucleotides), the $x_i$ succession element was:

$$x_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{i7}, x_{i8}).$$

And for the white point (one nucleotide) within the tree representation, the element $x_j$ was:

$$x_j = (x_{j1}, \ldots, x_{j4}, 0, 0, 0, 0),$$

for each succession element, we took the weighting factors from Galindo et al. [33], thus:

$$x_i = (PC_{1k}, PC_{2k}, PC_{3k}, PC_{4k}, PC_{1l}, PC_{2l}, PC_{3l}, PC_{4l}),$$

where $k$ and $l$ are the nucleotides forming a pair. In the case of a single nucleotide, the assignation would be:

$$x_i = (PC_{1k}, PC_{2k}, PC_{3k}, PC_{4k}, 0, 0, 0, 0).$$

Each secondary structure (Figure 3) would have an 8-tuple associated succession, i.e., for each RNA secondary structure we associated a succession whose size depended on the size of the RNA molecule. If a family of molecules was taken into account, then it would be possible to associate a succession set as $T = \{\bar{x}_1, \bar{x}_2, \dots\}$ where $\overline{x_j} = x_1, x_2, \dots, x_i, \dots, x_n$ and each element within the succession would correspond to an 8-tuple of positive real numbers.

## 5 Metrics as similarity measurements

One of the main points in the study of molecular reactivity is making comparisons amongst molecules to establish chemical similarity. Thus, if the molecular structure is represented as a succession, then molecules can be compared by analyzing these mathematical objects. We, thus, propose two new metrics, which are in turn compared to a standard one.

Since each term of the succession corresponds an 8-tuple of positive values, it is possible to associate a quantity between two 8-tuples, thus giving the possibility of capturing differences between them. For this purpose, we defined the following inner product:

**Definition 1.** Let $x_i$, $x_j$ be elements of either one or two different successions. The inner product $\langle x_i | x_j \rangle$ between these elements is defined as:

$$\langle x_i | x_j \rangle := \sum_{k=1}^{8} x_{ik} x_{jk} \ ,$$

which is normalized to $[0,1]$ by taking the maximum inner product value and dividing each inner product by it.

Now, with the succession representation obtained from the graph associated with the secondary RNA structure, the comparison or molecular similarity problem is addressed.

**Proposition 1.** *Let $\bar{x} = (x_i)_{i=1}^{n}$ and $\bar{y} = (y_i)_{i=1}^{n}$ be two successions in $T$, then the map $d_{suc} : T \times T \to \mathbb{R}$ defined as:*

$$d_{suc}(\bar{x}, \bar{y}) := \begin{cases} \frac{1}{k + \langle x_k | y_k \rangle} \ , & \text{if } \bar{x} \neq \bar{y}, \\ 0, & \text{if } \bar{x} = \bar{y}, \end{cases}$$

*where $k \in \{1, 2, \dots, n\}$ is the ordinal of the first term in which the successions differ and $\langle x_k | x_k \rangle$ is the $x_k$ and $y_k$ 8-tuples' inner product, then $d_{suc}$ is a metric over $T$.*

*Proof.* *i.* $d_{suc}(\overline{x},\overline{y}) \geqslant 0$.

Given that $\langle x_k|y_k\rangle \geqslant 0$ and $k > 0$, then

$$d_{suc}(\overline{x},\overline{y}) = \frac{1}{k+\langle x_k|y_k\rangle} > 0.$$

When $\overline{x} = \overline{y}$, $d_{suc}(\overline{x},\overline{y})$ is equal to zero by definition.

*ii.* $d_{suc}(\overline{x},\overline{y}) = 0 \iff \overline{x} = \overline{y}$. By definition.

*iii.* $d_{suc}(\overline{x},\overline{y}) = d_{suc}(\overline{y},\overline{x})$ for each $\overline{x}$, $\overline{y} \in T$ .

Given that the product between two real numbers is commutative, then $x_{ki} \cdot y_{ki} = y_{ki} \cdot x_{ki}$ and thus:

$$\langle x_k|y_k\rangle = \sum_{i=1}^{8} x_{ki} \cdot y_{ki} = \sum_{i=1}^{8} y_{ki} \cdot x_{ki} = \langle y_k|x_k\rangle,$$

therefore,

$$d_{suc}(\overline{x},\overline{y}) = \frac{1}{k+\langle x_k|y_k\rangle} = \frac{1}{k+\langle y_k|x_k\rangle} = d_{suc}(\overline{y},\overline{x}).$$

*iv.* Triangle inequality $d_{suc}(\overline{x},\overline{y}) \leqslant d_{suc}(\overline{x},\overline{z}) + d_{suc}(\overline{z},\overline{y})$ .

It is possible to assume without losing generality, that $d_{suc}(\overline{x},\overline{z}) = \frac{1}{r+\langle x_r|z_r\rangle}$, $d_{suc}(\overline{z},\overline{y}) = \frac{1}{q+\langle z_q|y_q\rangle}$ and that $\frac{1}{r+\langle x_r|z_r\rangle} \leqslant \frac{1}{q+\langle z_q|y_q\rangle}$. Given this initial condition, then

$$q + \langle z_q|y_q\rangle \leqslant r + \langle x_r|z_r\rangle.$$

(a) Case 1. If $q + \langle z_q|y_q\rangle < r + \langle x_r|z_r\rangle$, then $q \leqslant r$ and we have $z_i = y_i = x_i$ for each $i < q$; therefore, $d_{suc}(\overline{x},\overline{y}) \leqslant \frac{1}{q+\langle x_q|y_q\rangle}$ and we need to prove that:

$$\frac{1}{q+\langle x_q|y_q\rangle} \leqslant \frac{1}{q+\langle x_q|z_q\rangle} + \frac{1}{r+\langle z_r|y_r\rangle}.$$

When $q < r$, it can be guaranteed that $x_i = z_i$ for each $i < r$, then $\langle z_i|y_i\rangle = \langle x_i|y_i\rangle$ for each $i < r$; and as $q < r$, then $\langle z_q|y_q\rangle = \langle x_q|y_q\rangle$, and thus:

$$\frac{1}{q+\langle x_q|y_q\rangle} \leqslant \frac{1}{r+\langle x_r|z_r\rangle} + \frac{1}{q+\langle x_q|y_q\rangle},$$

which is an inequality that is always satisfied.

When $q = r$, we have $\frac{1}{q+1} \leqslant \frac{1}{q+\langle x_q|y_q\rangle} \leqslant \frac{1}{q}$ given that $\langle x_q|y_q\rangle \in [0,1]$. And the same holds true for $\frac{1}{q+\langle x_q|z_q\rangle}$ and $\frac{1}{q+\langle z_q|y_q\rangle}$. Assuming that $\frac{1}{q+\langle x_q|y_q\rangle} = \frac{1}{q}$ (the maximum possible value) and that $\frac{1}{q+\langle x_q|z_q\rangle}$ as well as $\frac{1}{q+\langle z_q|y_q\rangle}$ take the minimum possible value ($\frac{1}{q+1}$), then:

$$\frac{1}{q} \leqslant \frac{1}{q+1} + \frac{1}{q+1},$$

which is always satisfied because of $q \geqslant 1$.

(b) Case 2. If $q + \langle z_q|y_q\rangle = r + \langle x_r|z_r\rangle$, then $q = r$, the same as in the previous case, except when $\langle z_q|y_q\rangle = 0$ and $\langle x_r|z_r\rangle = 1$ or vice-versa.

When $\langle x_r|z_r\rangle = 1$ and $\langle z_q|y_q\rangle = 0$, we have $r+1 = q$, therefore $x_i = z_i$ and $y_i = z_i$ for each $i < r$. Guaranteeing $x_i = y_i$ for each $i < r$, thus $d_{suc}(\bar{x}, \bar{y}) \leqslant \frac{1}{r+\langle x_r|y_r\rangle}$. Therefore, we need to prove that:

$$\frac{1}{r + \langle x_r|y_r\rangle} \leqslant \frac{1}{q} + \frac{1}{r+1}.$$

As $y_i = z_i$ for each $i < q$, we have $\langle z_i|x_i\rangle = \langle x_i|y_i\rangle$ for each $i < q$, and since $r < q$, $\langle x_r|y_r\rangle = \langle x_r|z_r\rangle = 1$ this leads to

$$\frac{1}{r+1} \leqslant \frac{1}{q} + \frac{1}{r+1}.$$

And since $q = r+1$ then

$$\frac{1}{r+1} \leqslant \frac{1}{r+1} + \frac{1}{r+1},$$

an inequality that is always satisfied when $r > -1$. In our case $r$ is either greater than or equal to 1.

□

To obtain a similarity measurement suitable for comparing two successions in all their extensions, the previous metric was modified by building a new metric denoted $M_{suc}(\bar{x}, \bar{y})$ as follows:

**Proposition 2.** *Let $\bar{x} = (x_i)_{i=1}^n$ and $\bar{y} = (y_i)_{i=1}^n$ be two successions in $T$, then the map $M_{suc} : T \times T \to \mathbb{R}$ built as:*

$$M_{suc}(\bar{x}, \bar{y}) := \begin{cases} \sum_i^n \frac{1}{i+\langle x_i|y_i\rangle} \delta_{x_i y_i}, & \text{if } \bar{x} \neq \bar{y}, \\ 0 & \text{if } \bar{x} = \bar{y}, \end{cases}$$

*where $\delta_{x_i y_i} = 0$ if $x_i = y_i$ and $\delta_{x_i y_i} = 1$ whenever $x_i \neq y_i$, $0 \leqslant i \leqslant n$ and $\langle x_i|y_i\rangle$ is the inner product, then $M_{suc}$ is a metric over $T$.*

*Proof.* *i.* $M_{suc}(\overline{x},\overline{y}) \geqslant 0$.

Since $\delta_{x_iy_i}$ only represents the elements that are different between both successions then the same conditions applied in the previous proof are maintained for each element in the addition, *i.e.*, each term is greater than zero. Therefore,

$$M_{suc}(\overline{x},\overline{y}) = \sum_{i}^{n} \frac{1}{i + \langle x_i|y_i\rangle} \delta_{x_iy_i} \geqslant 0.$$

*ii.* $M_{suc}(\overline{x},\overline{y}) = 0$ if and only if $\overline{x} = \overline{y}$. By definition.

*iii.* $M_{suc}(\overline{x},\overline{y}) = M_{suc}(\overline{y},\overline{x})$ for every $\overline{x},\ \overline{y} \in T$.

From the previous metric, we have $\frac{1}{p+\langle x_p|y_p\rangle} = \frac{1}{p+\langle y_p|x_p\rangle}$, then adding these terms is symmetrical, which leads to:

$$M_{suc}(\overline{x},\overline{y}) = \sum_{i}^{n} \frac{1}{i + \langle x_i|y_i\rangle} \delta_{x_iy_i} = \sum_{i}^{n} \frac{1}{i + \langle y_i|x_i\rangle} \delta_{y_ix_i} = M_{suc}(\overline{y},\overline{x}).$$

*iv.* Triangle inequality $M_{suc}(\overline{x},\overline{y}) \leqslant M_{suc}(\overline{x},\overline{z}) + M_{suc}(\overline{z},\overline{y})$.

To prove this inequality, we will show that if one term is in $M_{suc}(\overline{x},\overline{y})$, then it will be either in $M_{suc}(\overline{x},\overline{z})$ or in $M_{suc}(\overline{z},\overline{y})$ or the three successions will differ in the same position, hence adding of these two terms is greater than a single term. However, if $\delta_{x_iy_i} = 1$ then either $\delta_{x_iz_i} = 1$ or $\delta_{z_iy_i} = 1$, because if $x_i \neq y_i$, then $x_i \neq z_i$ or $z_i \neq y_i$. Similarly, if $x_i = z_i$ and $z_i = y_i$ then $x_i = y_i$.

Supposing that $\delta_{x_iz_i} = 1$ and $\delta_{z_iy_i} = 0$, the opposite case is equivalent by symmetry. As $\delta_{z_iy_i} = 0$ then $z_i = y_i$, which leads to

$$\frac{1}{i + \langle x_i|z_i\rangle} = \frac{1}{i + \langle x_i|y_i\rangle},$$

and consequently, the term in which $\overline{x}$ and $\overline{y}$ differ is equal to one of the terms in which $\overline{x}$ and $\overline{z}$ differ.

On the other hand, when $\delta_{x_iz_i} = 1$ and $\delta_{z_iy_i} = 1$, we obtain:

$$\frac{1}{i + \langle x_i|y_i\rangle} \leqslant \frac{1}{i + \langle x_i|z_i\rangle} + \frac{1}{i + \langle z_i|y_i\rangle},$$

As it was proved in the previous metric, this inequality is satisfied for all ordinal $i$-values and since it is an addition of terms, the triangular inequality is also satisfied by the addition.

□

The last metric option is analogous to an Euclidean metric.

**Proposition 3.** *Let* $\overline{x} = (x_i)_{i=1}^{n}$ *and* $\overline{y} = (y_i)_{i=1}^{n}$ *be two successions in* $T$*, the map* $N_{suc} : T \times T \rightarrow \mathbb{R}$ *defines as:*

$$N_{suc}(\overline{x},\overline{y}) = \left( \sum_{k}^{n} \sum_{i=1}^{8} (x_{ki} - y_{ki})^2 \right)^{1/2},$$

*is a metric.*

*Proof.* This proof is the same as the proof for a Euclidean metric in a $\mathbb{R}^{8n}$ space.

□

## 6 Chemical Space

In the previous section, we approached comparing molecules by using metrics as similarity measures, allowing us to quantify such similarity.

The concept of chemical structure acquires its importance only when molecules are compared as a whole set. We consider that chemical structure is an emergent property manifested as the possibility of establishing relationships of similarity, order, complementariness and reactivity among molecules [34, 35]. Thus, the chemical structure concept allows us to establish classifications and/or equivalence relationships representing these kinds of relationships [12].

Molecular classifications use to be done by mean of clustering techniques that depends on the algorithm employed to cluster, here we propose an alternative well founded on the properties of the space defined by the representation of the molecules and the metric to compare them. Thus, we have codified the molecular secondary structure as mathematical objects within a set $T$, and a metric over this set. The next stage consists of exploring the $(T,d)$ space that could be called the *chemical space*. For such consideration, we can classify the set $T$ into classes using pretopologies.

The basic pretopological definitions found in the next subsection, show how a pretopological structure can be given to each metric space obtained in the previous section. These new spaces will be called *pretopological chemical spaces*.

## 6.1 Pretopological spaces

Being $T$ a set of molecular secondary structures represented as mathematical objects and $\mathscr{P}(T)$ the power set of $T$, it is possible to build a pretopological space as follows:

**Definition 2.** A pseudoclosure is a map $a(\_)$ from $:\mathscr{P}(T) \to \mathscr{P}(T)$ satisfying the following two conditions:

*1.* $a(\emptyset) = \emptyset$.

*2.* For each $A \in \mathscr{P}(T)$, $A \subseteq a(A)$.

A pretopological space is a pair $(T, a(\_))$ where $T$ is endowed with pseudoclosure $a(\_)$.

Subset $a(A)$ is also called pseudoclosure of $A$. The pseudoclosure of a set may not be idempotent.

**Definition 3.** Let $(T, a(\_))$ be a pretopological space. $(T, a(\_))$ is a *D*-pretopological space if the application $a(\_)$ satisfies not only *1.* and *2.* but the following as well:

*3.* For all $A, B \in \mathscr{P}(T)$, $\quad a(A \cup B) = a(A) \cup a(B)$.

**Definition 4.** A *V*-pretopological space $(T, a(\_))$ is a pretopological space that satisfies *4.*:

*4.* For all $A, B \in \mathscr{P}(T)$, $\quad A \subseteq B \Rightarrow a(A) \subseteq a(B)$.

**Lemma 1.** *A* D-*pretopological space is necessarily a* V-*pretopological space but the converse property is false (*$3. \Rightarrow 4.$*).*

**Definition 5.** A subset $F$ of $E$ such as $a(F) = F$ is called a closed subset of $T$ for $a(\_)$.

Let $\Im(T, a(\_))$ be the family of closed subsets of $T$ for $a(\_)$:
$\Im(T, a(\_)) = \{F \subseteq T : a(F) = F\}$ and $\Im(T, a(\_))^* = \Im(T, a(\_)) - \{\emptyset\}$.

*Property* 1. The intersection of closed subsets is a closed subset in a *V*-pretopological space.

**Definition 6.** Let $P$ be a subset of $T$. The closure of $F$ is the smallest closed subset in terms of inclusion in family $\Im(T, a(\_))$ containing $P$.

A closed subset noted as $F_x$ is the closure of a single element set $\{x\}$ of $T$.

*Property* 2. In a type *V*-pretopological space, each subset of $T$ has a closure.

*Property* 3. Two distinct elementary closed subsets $F_x$ and $F_y$ are either disjoint ($F_x \bigcap F_y = \emptyset$) or contain a non-empty intersection so that for all $F_z \in F_x \bigcap F_y$ we have $F_z \subseteq F_x \bigcap F_y$.

**Definition 7.** A minimal closed subset of $T$ regarding $a(\_)$ is an element of $\Im(T,a(\_))^*$ minimal in terms of inclusion in $\Im(T,a(\_))^*$.
Let $\Im_m(T,a(\_))$ represent the set of minimal closed subsets of $T$: $\Im_m(T,a(\_)) = \{F \in \Im(T,a(\_))^*, \neg(G \in \Im(T,a(\_))^* - \{F\}, G \subset F)\}$

The proofs and definitions presented in this section come from Bonnevay et al. [36]; Largeron and Bonnevay [37].

## 6.2 Pretopological space of the chemical structure

The classifications and comparisons of molecule set $T$ come from pretopologies in molecular space $(T,d)$. To do this, it is necessary to build $a(\_)$ and one way is:

**Proposition 4.** *Let $r$ be a real number. For each element $\bar{x} \in T$, we define $B_r(\bar{x}) = \{\bar{y} \in T : d(\bar{x},\bar{y}) \leqslant r\}$,* i.e., *a ball with center $\bar{x}$ and radius $r$. Given $B_r(\bar{x})$, the map $a_r(\_)$ is:*

$$a_r(A) =: \{\bar{x} \in T | B_r(\bar{x}) \cap A \neq \emptyset\}.$$

*And $a_r(\_)$ is a pseudoclosure over $T$.*

*Proof.*

1. $a_r(\emptyset) = \emptyset$.
   In the case of $A = \emptyset$, the intersection of $B_r(\bar{x}) \cap \emptyset = \emptyset$, for all $\bar{x} \in T$, since at least $\bar{x} \in B_r(\bar{x})$, therefore, $a_r(\emptyset) = \emptyset$.

2. For every $A \in \mathscr{P}(T)$, $A \subseteq a_r(A)$.
   For all $\bar{x} \in A$, $\bar{x} \in B_r(\bar{x})$ and therefore $x \in B_r(\bar{x}) \cap A$. Then, at least every $\bar{x} \in A$ belongs to $a_r(A)$.

□

By construction, each $r$-value has a possible pretopological space. Therefore, it is necessary to choose a criteria for assigning values to $r$ and then generating families of pretopological chemical spaces. To avoid choosing a subjective criteria, all values of $r$ must be analyzed to get all possible pretopological spaces (complete spectrum).

Since the number of elements in $T$ is finite, the number of unitary subsets is also finite, and thus it is possible to affirm that the number of different pretopological spaces is also finite, although the radius takes real positive values.

It is possible to get all the different pretopological spaces by increasing the radius. Therefore, when $r$ is very small a pretopological space analogous to the discrete topological space, where each unitary subset is a closed subset, will be obtained. For the case of a large value of $r$, the pretopological space with $T$ and $\emptyset$ as the only closed subsets will be obtained. These two cases correspond to the extreme cases.

One of the advantages of the use of pretopological spaces over topological structures corresponds to the possibility of generating different intermediate spaces by modulating $r$. The topological structure of the molecular space is finite and, with the metrics defined in this work, corresponds to discrete topological spaces without relevance to compared molecules.

Each of the closed sets defines an equivalence class in the pretopological spaces, where an equivalence class could be defined as: Let $X$ be a set, then each element in $X$ is an element class representation, i.e., only one element is necessary to characterize or describe the behavior of the elements within this set.

The intermediate pretopological spaces and the evolution of the equivalence classes are going to be of great importance, since the modulation of $r$ makes that one class link to another class in order to get a new equivalence class by increasing its value.
Due to the direct dependence of the different pretopological spaces with parameter $r$, this parameter can be understood as a dynamic variable. In the appearance or fusion of a closed set, it could be similar to either an evolutionary process or a molecular transformation. It could also be set as a resolution measurement regarding the degree for establishing relationships without an *a priori* definition of the process.

# 7 Application: Graph Directionality

Our previous model [14] uses elements that can be understood as local, due to their nature. As a consequence, this model is not able to encode all the possible differences of RNA chemical structures; for instance, in a particular nucleotide mutation where the first neighbors have a similar structure the model does not distinguish between them. This example is presented in Figure 4: a fragment of RNA (RNA **0**) is mutated by substituting G-C pair by a U-A pair in the positions 3 and 17 for RNA **A**, and 4 and 16 for RNA **B**, generating two slightly different RNAs. These differences are based on the fact that RNA is not completely symmetric and it also has a directionality which gives a particular behavior in these two different positions inside of the molecule.

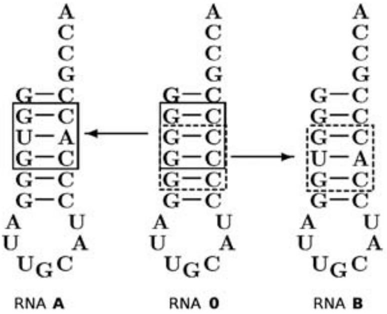

Figure 4: The mutation of RNA **0** in two different positions gives RNA **A** and RNA **B**.

To compare these three RNAs is possible to employ the graph-theoretical indices defined in Galindo et al. [14],

- **Randić index** $\chi^{\lambda}$.

$$\chi^{\lambda}_{a,c,d} = \sum_{paths} (B^{i}_{a,c,d} \times \cdots \times B^{k}_{a,c,d})^{1/2},$$

  where $\lambda$ corresponds to the path length (a value between 0 to 4) and $B_{a,c,d}$ is the modified valence.

- **Charge-valence index** $\mu$.

$$\mu = \frac{\sum_{n=1}^{N} \sum_{i=1}^{4} \nu_n \cdot q_{i,n}}{V}$$

  where $\nu$ is the standard graph-theoretical valence, $V = \sum_n \nu_n$, $N$ is the total number of vertices and $q_i$ is the quantum weight vector component associated to a particular vertex.

- **Sum of areas** $\sigma$.

$$\sigma = \sum_{i}^{n} \mathbb{D}^2_{i,j}$$

  where $\mathbb{D}^2$ is the area generated by two $i$ and $j$ vectors, which correspond to the weight factors of two adjacent nucleotides.

However, the decomposition analysis in motifs of 3 to 6 nucleotides, which were used in our initial model [14], causes RNA **A** and RNA **B** to have the same set of fragments. Therefore, the graph-theoretical indices, depending on the first neighbors, will be the same or very similar (see Table 1).

Within the first neighbors approach it is possible to affirm that both molecules are the same,

Table 1: Graph theoretical indices calculated for the three RNA molecules in Figure 4. The definition of these indices can be found in Galindo et al. [14].

| Index | RNA **0** | RNA **A** | RNA **B** | Index | RNA **0** | RNA **A** | RNA **B** |
|---|---|---|---|---|---|---|---|
| $\mu$ | 0.41951 | 0.44979 | 0.44979 | $\chi_c^0$ | 83.2876 | 87.5807 | 87.5807 |
| $\sigma$ | 21.1581 | 29.0294 | 29.0581 | $\chi_c^1$ | 87.1268 | 91.7395 | 91.7408 |
| $\chi_a^0$ | 12.8741 | 12.9952 | 12.9952 | $\chi_d^0$ | 18.6832 | 18.6790 | 18.6790 |
| $\chi_a^1$ | 8.2400 | 8.4109 | 8.4109 | $\chi_d^1$ | 18.5236 | 18.5132 | 18.5113 |
| $\chi_a^2$ | 6.4774 | 6.7302 | 6.7321 | $\chi_d^2$ | 23.7577 | 23.7323 | 23.7312 |
| $\chi_a^3$ | 6.7302 | 5.7314 | 5.7458 | $\chi_d^3$ | 32.6468 | 32.5937 | 32.5959 |
| $\chi_a^4$ | 6.7321 | 4.4024 | 4.3724 | $\chi_d^4$ | 39.2112 | 39.1289 | 39.1377 |

Table 2: Molecular euclidean distance matrix using the vector built with the 14 graph theoretical indices.

| | RNA 0 | RNA A | RNA B |
|---|---|---|---|
| RNA 0 | 0.000 | | |
| RNA A | 10.402 | 0.000 | |
| RNA B | 10.430 | 0.045 | 0.000 |

which is false, because it is well known that mutations in different positions give the molecule particular characteristics, and sometimes these mutations can cause a partial or total loss of the biological activity. Thus, although we can consider graph theoretical indices with higher molecular environments, i.e. including second and four neighbors, the values for the system will change from thousandths to tenths and this change will not be enough to establish structure activity relationships. Furthermore, the numbers are similar enough to make the molecules almost indistinguishable, implying a high similitude between them.

This phenomenon is shown if we use the graph theoretical indices set as a vector to represent the molecule and the euclidean distance as a similarity measure. It is possible to observe in Table 2, that the distance between the RNA **0** molecule and the two mutated molecules is of the order of tens, while the distance between RNA **A** and **B** is 0.045. Therefore, the latter has negligible value in comparison to the former, leading to conclude that RNA **A** and RNA **B** are the same.

The advantage of using the succession representation, that comes from the tree representation, is to recover the digraph character intrinsic in RNA molecules. In Table 3 the 8-tuples for every succession element are shown. Every term in the tuples comes from a quantum weight factor based on the Principal Component (PC) analysis of the partial atomic charges, [33] and, as it was mentioned before, the 8-tuple representing a hydrogen bond pair of nucleotides will have four PCs coming from every nucleotide, while the 8-tuple for a single strand nucleotide will have four components that are zero.

Table 3: 8-Tuples of positive real numbers for every succession element of the two mutated RNA examples.

| RNA **A** | RNA **B** | Motif | $PC_{1k}$ | $PC_{2k}$ | $PC_{3k}$ | $PC_{4k}$ | $PC_{1l}$ | $PC_{2l}$ | $PC_{3l}$ | $PC_{4l}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | $x_1$ | 5'- ***G***G-3'<br>3'-G***C***C-5' | 1.24010 | 1.05118 | 0.41770 | 0.39401 | 1.04050 | 1.55828 | 0.13912 | 0.75109 |
| $x_2$ | $x_3$ | 5'-G***G***U-3'<br>3'-C***C***A-5' | 1.19733 | 0.94704 | 0.32115 | 0.29314 | 1.06064 | 1.60006 | 0.23383 | 0.64446 |
| $x_3$ | $x_4$ | 5'-G***U***G-3'<br>3'-C***A***C-5' | 0.30197 | 1.13813 | 0.22830 | 0.72278 | 1.37867 | 1.16706 | 0.34749 | 0.68019 |
| $x_4$ | $x_5$ | 5'-U***G***G-3'<br>3'-A***C***C-5' | 1.26737 | 1.05100 | 0.38354 | 0.38226 | 1.03928 | 1.62765 | 0.24149 | 0.65505 |
| $x_5$ | $x_2$ | 5'-G***G***G-3'<br>3'-C***C***C-5' | 1.21895 | 0.99706 | 0.36996 | 0.35165 | 1.06227 | 1.60546 | 0.24010 | 0.65335 |
| $x_6$ | $x_6$ | 5'-G***G***A-3'<br>3'-CC -5' | 1.06525 | 1.59848 | 0.24748 | 0.66327 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| $x_7$ | $x_7$ | 5'-G***A***U-3'<br>3'-C -5' | 0.22471 | 1.04005 | 0.25452 | 0.57245 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| $x_8$ | $x_8$ | 5'-A***U***U-3' | 1.34780 | 1.08047 | 0.28851 | 0.59910 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| $x_9$ | $x_9$ | 5'-U***U***G-3' | 1.36323 | 1.09136 | 0.30995 | 0.61224 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| $x_{10}$ | $x_{10}$ | 5'-U***G***C-3' | 1.04245 | 1.50557 | 0.19985 | 0.58359 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| $x_{11}$ | $x_{11}$ | 5'-A***C***G-3' | 1.07897 | 0.86268 | 0.22218 | 0.45961 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| $x_{12}$ | $x_{12}$ | 5'-C***A***U-3' | 0.20933 | 1.08871 | 0.30624 | 0.52344 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| $x_{13}$ | $x_{13}$ | 5'-C -3'<br>3'-C***U***A-5' | 1.33970 | 1.07317 | 0.23220 | 0.64923 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| $x_{14}$ | $x_{14}$ | 5'- G-3'<br>3'-C***G***C-5' | 1.03067 | 1.52105 | 0.14538 | 0.64691 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| $x_{15}$ | $x_{15}$ | 5'-G***C***C-3' | 1.06026 | 0.82388 | 0.19157 | 0.42897 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| $x_{16}$ | $x_{16}$ | 5'-C***C***A-3' | 1.11606 | 0.89724 | 0.27969 | 0.52429 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| $x_{17}$ | $x_{17}$ | 5'-CC***A***-3' | 10.24648 | 1.11871 | 0.34481 | 0.57572 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |

Having defined the successions we applied the similarity measures defined in Section 3. The results are presented in Table 4. In contrast to the case of graph theoretical indices, the succession representation makes the RNA **A** and RNA **B** distances be different than zero. Hence, we can conclude that the molecules are different. Analyzing values from $d_{suc}$ is possible to see that the difference among the three molecules is very small, and the distance between RNA **A** and **B** is similar to the distance between RNA **0** and RNA **A**. From this point of view, this metric encodes information about the first secondary structure mutation in the RNA **0** molecule: if the first terms of the succession are the same, i.e., the mutation is made in a far nucleotide, the more similar the molecules will be. This statement will be correct if we consider the tRNAs case, where one of the most important nucleotides corresponds to the discriminator base and its environment [38,39].

Table 4: Distance measure results over the three different RNAs

| | $d_{suc}$ | $M_{suc}$ | $N_{suc}$ |
|---|---|---|---|
| (RNA **A**, RNA **B**) | 0.272 | 0.484 | 1.659 |
| (RNA **A**, RNA **0**) | 0.270 | 0.2702 | 1.150 |
| (RNA **B**, RNA **0**) | 0.213 | 0.213 | 1.150 |

The $M_{suc}$ metric, besides encoding the same information as $d_{suc}$, also includes information

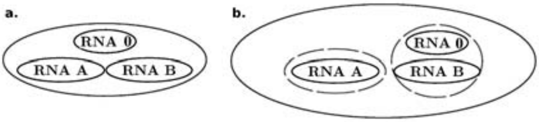

Figure 5: Pretopological spaces generated by the different metrics, changing the variational parameter $r$. **a.** Two pretopological spaces obtained using the $N_{suc}$ metric. **b.** Three pretopological spaces obtained for the $M_{suc}$ metric as well as the $d_{suc}$ metric.

about all the differences between the molecules, being more appropriate when a comparison between systems is done. Comparing RNA **A** and **B**, we expect to have a higher distance, since they differ in two positions, than each of them with RNA **0**. In this sense, the $M_{suc}$ distance is more appropriate because the difference between them is two times higher than difference of each one with respect to the non-mutated RNA.

Finally, the $N_{suc}$ metric shows that the comparison done nucleotide by nucleotide could be problematic because the value of the distance between RNA **A** and RNA **0**; and RNA **B** and RNA **0** is exactly the same and it is pretty similar to the value for the two mutated RNAs. Therefore, $N_{suc}$ is unable to encode the directionality of RNA molecules, i.e. every nucleotide has the same weight no matters their position in the sequence.

We would like to point out that, unlike the graph theoretical indices, the metric measures proposed do not consider RNA **A** and RNA **B** to be the same. Moreover, the values obtained are consistent with the observation that the mutated RNAs are more similar to RNA **0** than each other. The proposed distances are able to encode the mutation's position and the directionality associated to RNA molecules.

Besides the differences among the metrics, we want to illustrate the use of pretopologies as a molecular classification tool, allowing the understanding of this process in a more general way. Considering the various pretopologies obtained for each metric due to the variational parameter $r$; it is possible to observe that in the case of the pretopologies associated to $N_{suc}$ the situation is not very rich from the structural point of view. We have an initial scenario, where the three molecules correspond to a closed set, i.e., every molecule is a different class. On the other hand, if the level of resolution decreases all molecules are in the same class (Figure 5a).

In the case of the other two metrics, the result is richer. When the value of $r$ is very small, every molecule corresponds to a closed set; however, by increasing the $r$ value it is possible to reorder the molecules in a new and different pretopological space, which has two equivalence classes: the pair $\{$RNA **0**, RNA **B**$\}$ and $\{$RNA **A**$\}$. Finally, increasing the parameter $r$ further, a condensed pretopological classification is observed where every molecule belongs to the same class (Figure 5b). The generation of this intermediate pretopological space is very interesting especially for the tRNA molecules. For instance, if we analyze RNA **A**, it is closer to a hypothetical discriminator base position, affecting the chemical environment in a more decisive way than for RNA **B**, and therefore its biological activity.

This simple example shows that using pretopological spaces as a classification tool enriches the comparison between molecules, extending the concept of similarity beyond a pair comparison. It also allows the classification in different levels of similarity, evidencing an evolutionary process.

## 8 Conclusion

This work proposes a new representation that is able to encode basic information about the directionality of RNA, as well as the chemical environment, since the weight factors for every element of the new succession arrive from statistical variables derived from quantum partial charges. Successions preserve the information associated with the connectivity of the molecules, i.e. encoding the secondary structure information of the RNA.

In the same way, the use of metrics as chemical similarity measures have shown to be a good option, since the results obtained correlated well with the similarity expected based on informal arguments. Finally, in this work we defined a mathematical space for representation and classification of RNA molecules, where the classification was performed using properties derived from pretopological spaces, resulting in different levels of similarity.

*Acknowledgments*: We want to acknowledge N. Di Russo, Davita Watkins, B. Miller III and J. Bueren-Calabuig for useful comments about this paper. We also want to acknowledge Dr. Clara I. Bermúdez for her helpful discussions and contributions to this paper, as well as some templates for the graphics.

## References

[1] P. Clote, Introduction to special issue on RNA, *J. Math. Biol.* **56** (2008) 3–13.

[2] C. Laing, T. Schlick, Computational approaches to RNA structure prediction, analysis and design, *Curr. Opin. Struc. Biol.* **21** (2011) 306–318.

[3] M. Shimizu, H. Asahara, K. Tamura, T. Hasegawa, H. Himeno, The role of anticodon bases and discriminator nucleotide in the recognition of *E. Coli tRNAs* by their aminoacyl-tRNA synthetases, *J. Mol. Evol.* **35** (1992) 436–443.

[4] A. R. Srinivasan, W. K. Olson, Molecular models of nucleic acid triple helixes. II. PNA and $2' - 5'$ backbone complexes, *J. Am. Chem. Soc.* **120** (1998) 492–499.

[5] D. L. Nelson, A. L. Lehninger, M. M. Cox, *Lehninger Principles of Biochemistry*, Freeman, New York, 2008.

[6] N. B Leontis, A. Lescoute, E. Westhof, The building blocks and motifs of RNA architecture, *Curr. Opin. Struc. Biol.* **16** (2006) 279–287.

[7] M. Bon, G. Vernizzi, H. Orland, A. Z., Topological classification of RNA structures, *J. Mol. Biol.* **379** (2008) 900–911.

[8] J. E. Andersen, R. C. Penner, C. M. Reidys, M. S. Waterman, Topological classification and enumeration of RNA structures by genus, *J. Math. Biol.* **67** (2013) 1261–1278.

[9] J. Villaveces, E. E. Daza, On the topological approach to the concept of chemical structure, *Int. J. Quantum Chem.* **24** (1990) 97–106.

[10] J. Villaveces, E. E. Daza, The concept of chemical structure, in: D. H. Rouvray (Ed.), *Concepts in Chemistry: A Contemporary Challenge*, Wiley, New York, 1997, pp. 101–132.

[11] C. Li, I. Xing, X. Wang, Analysis of similarity of RNA secondary structures based on a 2d graphical representation, *Chem. Phys. Lett.* **458** (2008) 249–252.

[12] A. Bernal, E. E. Daza, On the epistemological and ontological status of chemical relations, *HYLE* **16** (2010) 80–103.

[13] D. R. Koessler, D. J Knisley, J. Knisley, T. Haynes, A predictive model for secondary RNA structure using graph theory and a neural network, *BMC Bioinformatics* **11** (2010) S21.

[14] J. F. Galindo, C. Bermúdez, E. E. Daza, tRNA structure from a graph and quantum theoretical perspective, *J. Theor. Biol.* **240** (2006) 574–582.

[15] F. Harary, *Graph Theory*, Addison–Wesley, Reading, 1969.

[16] A. T. Balaban, Applications of graph theory in chemistry, *J. Chem. Inf. Comput. Sci.* **25** (1985) 334–343.

[17] F. Bai , D. Li, T. Wang, A new mapping rule for RNA secondary structures with its applications, *J. Math. Chem.* **43** (2008) 932–943.

[18] A. Churkin, I. Gabdank, D. Barash, On topological indices for small RNA graphs, *Comput. Biol. Chem.* **41** (2012) 35–40.

[19] N. Kim, L. Petingi, T. Schlick, Network theory tools for RNA modeling, *WSEAS Transact. Math.* **12** (2013) 941–955.

[20] M. S. Waterman, Secondary structure of single–stranded nucleic acids, *Adv. Math. Suppl. Stud.* **1** (1978) 167–212.

[21] C. Reidys, P. F. Stadler, Biomolecular shapes and algebraic structures, *Comput. Chem* **20** (1996) 85–94.

[22] F. Rosselló, Reidy's and Stadler's metrics for RNA contact structures, *Math. and Comput. Modelling* **40** (2004), 771–776.

[23] S. Le, R. Nussinov, J. Maizel, Tree graphs of RNA secondary structures and their comparisons, *Comput. Biomed. Res.* **22** (1989) 461–473.

[24] G. Benedetti, S. Moroseti, A graph–topological approach to recognition and similarity in RNA secondary structures, *Biophys. Chem.* **59** (1996) 179–184.

[25] Z. Mihalić, N. Trinajstić, A graph–theoretical approach to structure–property relationships (SYM), *J. Chem. Educ.* **69** (2002) 701–712.

[26] H. H. Gan, S. Pasquali, T. Schlick, Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design, *Nucleic Acids Res.* **31** (2003) 2926–2943.

[27] P. Schuster, W. Fontana, P. F. Stadler, I. L. Hofacker, From sequences to shapes and back: a case study in RNA secondary structures, *Proc. R. Soc. Lond.* **255** (1994) 279–284.

[28] M. Zuker, The use of dynamic programming algorithms in RNA secondary structure prediction, in: M. S. Waterman (Ed.), *Mathematical Methods for DNA Sequences*, CRC Press, Boca Raton, 1989, pp. 159–184.

[29] M. A. Steel, D. Penny, Distributions of tree comparison metrics: Some new results, *Systematic Biol.* **42** (1993) 126–141.

[30] V. Moulton, M. Zuker, M. Steel, R. Pointon, D. Penny, Metrics on RNA secondary structures, *J. Comput. Biol.* **7** (2000) 277–292.

[31] J. J. Nieto, A. Torres, M. M. Vázquez–Trasande, A metric space to study differences between polynucleotides, *Appl. Math. Lett.* **16** (2003) 1289–1294.

[32] M. Llabrés, F. Rosselló, A new of metric for biopolymer contact structures, *Comput. Biol. Chem.* **28** (2004) 21–37.

[33] J. F. Galindo, C. Bermúdez, E. E. Daza, A classification of central nucleotides induced by the influence of neighboring nucleotides in triplets, *J. Mol. Struct. (Theochem)* **769** (2006) 103–109.

[34] E. Daza, *Un concepto más flexible de estructura química basado en las variables espaciales y de carga asociadas con los núcleos*, Ph.D. thesis, Universidad Nacional de Colombia, 2000.

[35] E. E. Daza, A. Bernal, Energy bounds for isoelectronic molecular sets and the implicated order, *J. Math. Chem.* **38** (2005) 247–263.

[36] S. Bonnevay, M. Ure, N. Largeron, N. Nicolayannis, A pretopological approach for structuring data in non–metric spaces, *El. Notes Discr. Math.* **2** (1999) 1–9.

[37] C. Largeron, S. Bonnevay, A pretopological approach for structural analysis, *Inf. Sci.* **144** (2002) 169–185.

[38] Y. M. Hou, Discriminating among the discriminator bases of tRNAs, *Chem. Biol.* **4** (1997) 93–96.

[39] L. Bonnefond, R. Giege, J. Rudinger–Thirion, Evolution of the tRNA(Tyr)/TyrRS aminoacylation systems, *Biochimie* **87** (2005) 873–883.