Comparing biomolecular contact structures: the algebraic way

Francesc Rosselló

Dept. of Mathematics and Computer Science,
Research Inst. of Health Science (IUNICS),
Univ. of the Balearic Islands, 07122 Palma (Spain)
E-mail: cesc.rossello@uib.es

Abstract

Contact structures are simplified models of RNA molecules, proteins, and other biopolymer three-dimensional structures. In this paper we study several algebraic representations of biomolecular contact structures and some metrics on sets of contact structures of a given length derived from these representations. We take the opportunity to recall a few problems that remain open in this field of research.

Keywords: biomolecular contact structure, RNA secondary structure, metric, distance, symmetric group, monomial ideal.

1 Introduction

A biopolymer can always be viewed as an oriented chain of monomers, which in its turn can be mathematically described as a word over a suitable alphabet. This word is called the *primary structure* of the molecule. For instance, an RNA molecule is a chain of (ribo)nucleotides, each one of them characterized by the base attached to it: adenine (A), cytosine (C), guanine (G), or uracil (U). Thus, the primary structure of an RNA molecule with n nucleotides is a word of length n over $\{A, C, G, U\}$. In a similar way, proteins are chains of aminoacids, and hence the primary structure of a protein is a word over a 20-letter alphabet, for instance

$$\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\},$$

each letter representing an aminoacid: A for Alanine, C for Cysteine, D for Aspartic acid, etcetera.

In the cell and *in vitro*, each RNA molecule and protein folds into a three-dimensional structure that determines its biochemical function. As different levels of graining are suitable for different problems [17, 20], we can sometimes forget about the detailed description of this three-dimensional structure and

consider only a simplified model of it, called its *contact structure*: the set of all pairs of non-consecutive monomers that are, in some specific sense, neighbors in the three-dimensional structure. Such a contact structure can be mathematically described as an undirected graph without multiple edges or self-loops, with sets of nodes representing the monomers numbered according to their position along the molecule's backbone, and such that there does not exist any edge between consecutive nodes. The *length* of such a contact structure is its number of nodes and its edges are consistently called *contacts*.

The secondary structures of RNA molecules form a special class of contact structures. In them, contacts represent the *hydrogen bonds* between pairs of bases that hold together the three-dimensional structure. A hydrogen bond can only form between bases that are at least four positions apart in the chain, but we shall not take this restriction into account here and we shall only forbid the existence of hydrogen bonds between consecutive bases. A restriction, called the *unique bonds condition*, is added to the definition of RNA secondary structure: a base can only pair with at most another base. It is usual to impose a further restriction on *RNA secondary structures*, by forbidding the existence of (*pseudo*)*knots*, i.e., of contacts that "cross" each other. This restriction has its origin in the first dynamic programming methods to predict RNA secondary structures [23, 27], but real RNA structures can contain knots, which are moreover important structural elements [24].

Unfortunately, beyond RNA secondary structures, the representation of the neighborhood in three-dimensional structures of RNA molecules and proteins needs contact structures without unique bonds. The full three-dimensional structure of an RNA molecule may contain sets of contacts that violate the unique bonds condition, like base triplets and guanine platforms [1, 15], and in the three-dimensional structure of a protein, one aminoacid can be spatial neighbor of several aminoacids [4, 7]. But even in this general case, the neighborhood of pairs of consecutive monomers is not reflected in the contact structure, because their spatial closeness is understood as a consequence of their position in the backbone.

An important problem in molecular biology is the comparison of the three-dimensional structures of RNA and protein molecules, because it is assumed that a preserved three-dimensional structure corresponds to a preserved function. In this paper we shall restrict ourselves to the measurement of the similarity of contact structures of a fixed length, which has an interest in itself. It is used for instance in the analysis of the ensemble of suboptimal solutions provided by a given algorithm to the problem of determining the secondary structure of a given RNA molecule [26, §IX]. It can also be used to compare the output of different prediction algorithms applied to the same RNA molecule or protein, to assess their performance (G. Valiente, private communication). This similarity measurement lies also at the basis of the study of the mapping that assigns to each RNA molecule or protein the structure it folds into [8, 19]. Finally, the simultaneous comparison of primary and contact structures of two biomolecules can be carried out by first finding a global alignment of the primary structures and then comparing the contact structures, that become of the same length after

adding to each one of them fake, isolated nodes to represent all empty positions in the alignment; we shall report on it elsewhere. Several metrics on the set of all contact structures with unique bonds of a given length have been introduced so far in the literature: for instance, Zuker's d_Z [25, 26] and *mountain* metrics [16], and distances based on tree and graph edition [10, 16, 21].

In this paper we analyze several metrics based on the representation of contact structures as elements of algebraic structures, or even as algebraic structures themselves. This line of research was initiated by C. Reidys and P. F. Stadler in a seminal paper on the algebraic representation of biomolecular structures [17], where they introduced three abstract metrics on the set of RNA contact structures of a fixed length derived from their algebraic models and independent of any notion of graph edition. Reidys and Stadler ended their paper by asking some relevant questions on their models and metrics, including the extension of their metrics to arbitrary contact structures. Their work has been recently revisited by members of our "Computational biology and bioinformatics" group at the IUNICS, and as a consequence several of their questions have been answered and some other questions have arisen. The main goal of this survey is to discuss Reidys and Stadler's models and metrics and their generalization to contact structures without unique bonds, and to attract the attention of the reader towards some interesting algebraic problems that remain open in this field. We also introduce some new ideas on ideal-based representations and metrics: §§4.1, 4.3 and 4.4 and part of §4.2 are presented here for the first time.

In our view, this is a field where the participation of algebrists could be invaluable in the development of new representations and metrics that can be used in the study and comparison of biomolecular contact structures. After all, and in words of Moulton, Zuker *et al*, "[...] generally speaking, it is probably safest to try as many metrics as possible" [16, p. 290].

2 Contact structures

In this section we formalize the notion of a contact structure of a biomolecule and its refinements as discussed in the Introduction. We also take the opportunity to introduce some conventions and notations that will be used henceforth, usually without any further notice.

Let $[n]$ denote from now on the set $\{1, \ldots, n\}$, for every positive integer n.

Definition 1 *A* contact structure *of* length n *is an undirected graph without multiple edges or self-loops* $\Gamma = ([n], Q)$*, for some* $n \geqslant 1$*, whose arcs* $\{j, k\} \in Q$*, called* contacts*, satisfy the following condition:*

(i) For every $j \in [n]$*,* $\{j, j+1\} \notin Q$*.*

A contact structure $\Gamma = ([n], Q)$ has unique bonds *when it satisfies the following further condition:*

(ii) For every $i \in [n]$*, if* $\{i, j\}, \{i, k\} \in Q$*, then* $j = k$*.*

An RNA secondary structure *is a contact structure with unique bonds* $\Gamma = ([n], Q)$ *that satisfies the following condition:*

(iii) If $\{i,j\},\{k,l\} \in Q$ *and* $i < k < j$*, then* $i < l < j$*.*
Let $\mathcal{C}_n$ *and* $\mathcal{U}_n$ *denote, respectively, the sets of all contact structures and of all contact structures with unique bonds of length* n*.*

As it can be seen, condition (i) translates the impossibility of a contact between two consecutive bases, condition (ii) translates the unique bonds condition, and condition (iii) forbids the existence of (pseudo)knots.

We shall denote a contact $\{j,k\}$ by $j\cdot k$ or $k\cdot j$, without distinction. A node $i \in [n]$ is said to be *isolated* in Γ when it is not involved in any contact. The *empty contact structure* of length n is $([n],\emptyset)$, i.e., the contact structure of length n with all its nodes isolated.

Given contact structures of the same length $\Gamma_1 = ([n],Q_1), \Gamma_2 = ([n],Q_2)$, their *union* and *symmetric difference* are, respectively, the contact structures

$$\Gamma_1 \cup \Gamma_2 = ([n], Q_1 \cup Q_2), \qquad \Gamma_1 \Delta \Gamma_2 = ([n], Q_1 \Delta Q_2).$$

Notice that, even if Γ_1, Γ_2 are RNA secondary structures, the contact structures $\Gamma_1 \cup \Gamma_2$ and $\Gamma_2 \Delta \Gamma_2$ need not satisfy the unique bonds condition.

From now on, and unless otherwise stated, given any contact structure Γ or Γ_i, $i = 1, 2, \ldots$, we shall always denote its set of contacts by Q or Q_i, respectively.

3 Group-based representations and metrics

3.1 Involution representation and metric

Two of the algebraic models of RNA secondary structures proposed by Reidys and Stadler [17] were based on the interpretation of a contact as a transposition of the nodes it pairs. In the model recalled in this subsection, a contact structure with unique bonds is represented then as the product of these transpositions.

More specifically, let S_n be the symmetric group of permutations of $[n]$, for every $n \geqslant 1$. Then, Reidys and Stadler associated to every $\Gamma = ([n],Q) \in \mathcal{U}_n$ the permutation

$$\pi(\Gamma) = \prod_{i\cdot j \in Q} (i,j) \in S_n$$

where each (i,j) denotes the transposition in S_n defined by $i \leftrightarrow j$.

The unique bonds condition implies that, for every $\Gamma \in \mathcal{U}_n$, $\pi(\Gamma)$ is well defined, because all transpositions in the product $\prod_{i\cdot j\in Q}(i,j)$ are pairwise disjoint and hence they commute with each other, and that it is an involution. They proved then that the mapping $\pi : \mathcal{U}_n \to S_n$ is injective, and they used this fact to define the following metric, which we shall call henceforth the *involution metric*.

Proposition 1 [17, Cor. 1] *The mapping* $d_{inv} : \mathcal{U}_n \times \mathcal{U}_n \to \mathbb{R}$ *that sends every* $(\Gamma_1,\Gamma_2) \in \mathcal{U}_n^2$ *to the least number* $d_{inv}(\Gamma_1,\Gamma_2)$ *of transpositions necessary to represent the permutation* $\pi(\Gamma_1)\pi(\Gamma_2)$*, is a metric.*

An explicit description of this metric was derived in [18]. To recall this description we need to introduce some notions that will also be used in other sections.

Given $\Gamma_1, \Gamma_2 \in \mathcal{U}_n$, the *orbits* of their symmetric difference $\Gamma_1 \Delta \Gamma_2$ are the connected components of this graph that have at least two nodes. In other words, the orbits of $\Gamma_1 \Delta \Gamma_2$ are the subsets $\{i_1, i_2, \ldots, i_m\} \subseteq [n]$ with $m \geqslant 2$, such that

$$i_1 \cdot i_2, i_2 \cdot i_3, \ldots, i_{m-1} \cdot i_m \in Q_1 \Delta Q_2$$

and maximal with this property, i.e., such that if there is some other contact in $Q_1 \Delta Q_2$ involving some element of this subset, this contact can only be $i_1 \cdot i_m$. The unique bonds condition implies that if $\{i_1, i_2, \ldots, i_m\}$ is such an orbit, then $i_1 \cdot i_2, i_3 \cdot i_4, \ldots$ belong to one of the sets Q_1 or Q_2 and $i_2 \cdot i_3, i_4 \cdot i_5, \ldots$ belong to the other one.

We shall say that an orbit is *closed* if $m \geqslant 3$ and $i_1 \cdot i_m \in Q_1 \cup Q_2$, and that it is *open* when it is not closed; see Fig. 1. The closed orbits are the non-trivial cyclic connected components of $\Gamma_1 \Delta \Gamma_2$. The unique bonds condition implies that the length of a closed orbit is always even (and then at least 4) and all its nodes have degree exactly 2. As far as the open orbits goes, it is straightforward to check that an open orbit can have any length (greater than 1) and that all nodes in it have degree 2 except two of them, which have degree 1.

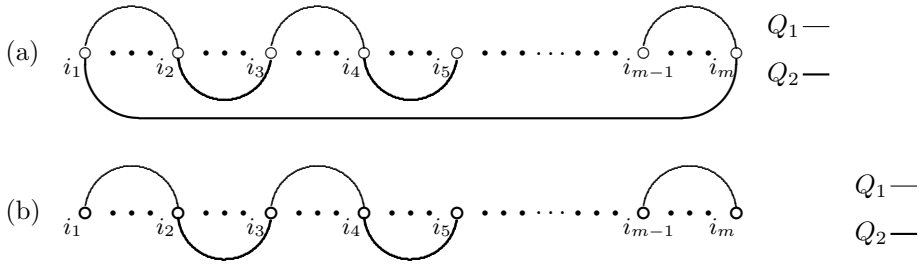

Figure 1: A closed (a) and an open (b) orbit of length m.

Let $\Theta(\Gamma_1, \Gamma_2)$ be the number of closed orbits of $\Gamma_1 \Delta \Gamma_2$.

Proposition 2 [18, Prop. 2] *For every* $\Gamma_1, \Gamma_2 \in \mathcal{U}_n$,

$$d_{inv}(\Gamma_1, \Gamma_2) = |Q_1 \Delta Q_2| - 2\Theta(\Gamma_1, \Gamma_2).$$

This characterization motivates us to study the behavior of $\Theta(\Gamma_1, \Gamma_2)$. For instance, we might ask ourselves what is the probability that, given two contact structures with unique bonds Γ_1, Γ_2 of a given length n, $\Theta(\Gamma_1, \Gamma_2) = 0$ and hence $d_{inv}(\Gamma_1, \Gamma_2) = |Q_1 \Delta Q_2|$. R. Alberich and J. Miró have carried out some preliminary numerical simulations with randomly generated RNA secondary structures of large length that show that this probability is high and it decreases with n, but the question is far from settled.

Open problem 1 *To determine the probability distribution of the values of* Θ *on pairs of contact structures with unique bonds (or even RNA secondary structures) of a given length.*

Can we generalize these involution representation and metric to arbitrary contact structures? A meaningful generalization of the involution representation using the same definition does not seem easy. Notice that the product of transpositions defining $\pi(\Gamma)$ is only well-defined if the transpositions appearing in it commute with each other. Therefore, this definition does not make sense for arbitrary contact structures unless some convention is specified on the order how these transpositions must be composed. And not every such convention will work if we want the resulting mapping $\pi : \mathcal{C}_n \to S_n$ to be injective. For instance, a natural convention would be to compose (from right to left) the transpositions in their lexicographic order:

$$\pi(\Gamma) = (i_t, j_t) \cdots (i_2, j_2)(i_1, j_1)$$

if

$$\Gamma = ([n], \{i_1 \cdot j_1, i_2 \cdot j_2, \ldots, i_t \cdot j_t\})$$

with $(i_1, j_1) < (i_2, j_2) < \cdots < (i_t, j_t)$ with respect to the order given by $(i, j) < (k, l)$ if and only if $i < k$ or $i = k$ and $j < l$. But then, taking for instance $\Gamma_1 = ([5], \{1{\cdot}3, 1{\cdot}5, 3{\cdot}5\})$ and $\Gamma_2 = ([5], \{1{\cdot}5\})$, we have that

$$\pi(\Gamma_1) = (3, 5)(1, 5)(1, 3) = (1, 5) = \pi(\Gamma_2).$$

On the positive side, and following Reidys and Stadler's original definition of the involution metric, we have that if a suitable, injective mapping $\pi : \mathcal{C}_n \to G$ is discovered, for some finite group G, then the mapping $d_\pi : \mathcal{C}_n \times \mathcal{C}_n \to \mathbb{R}$ sending every pair of contact structures (Γ, Γ') to the least number of some fixed type of generators of G necessary to represent $\pi(\Gamma')^{-1}\pi(\Gamma)$, will be a metric on $\mathcal{C}_n$.

In all, the following problem remains open:

Open problem 2 *To generalize the involution representation and metric to the whole* $\mathcal{C}_n$*.*

We shall propose a possible path leading to the solution of this problem in §3.3.

3.2 Subgroup representation and metric

In their second representation of RNA secondary structures based on the interpretation of contacts as transpositions, Reidys and Stadler associated to every $\Gamma \in \mathcal{U}_n$ the subgroup $G(\Gamma)$ of S_n generated by the set of the transpositions corresponding to the contacts in Γ:

$$G(\Gamma) = \langle \{(i, j) \mid i \cdot j \in Q\} \rangle.$$

They proved that the mapping $\Gamma \mapsto G(\Gamma)$ is an embedding of $\mathcal{U}_n$ into the set $\mathrm{Sub}(S_n)$ of subgroups of S_n, and they used this embedding to define the following metric, which we shall call henceforth the *subgroup metric*.

Proposition 3 **[17, §5.2]** *The mapping* $d_{sgr} : \mathcal{U}_n \times \mathcal{U}_n \to \mathbb{R}$ *defined by*

$$d_{sgr}(\Gamma_1, \Gamma_2) = \ln \left| \frac{G(\Gamma_1) \cdot G(\Gamma_2)}{G(\Gamma_1) \cap G(\Gamma_2)} \right|$$

is a metric.

Such a complicated formula is actually equivalent to a very simple measure: the cardinal of the symmetric difference of the sets of contacts.

Proposition 4 **[18, Prop. 4]** *For every* $\Gamma_1, \Gamma_2 \in \mathcal{U}_n$,

$$d_{sgr}(\Gamma_1, \Gamma_2) = (\ln 2)|Q_1 \Delta Q_2|.$$

In particular, had Reidys and Stadler used $\log_2$ instead of ln, their definition of $d_{sgr}(\Gamma_1, \Gamma_2)$ would coincide with $|Q_1 \Delta Q_2|$.

What about the extension of these subgroup representation and metric to the whole $\mathcal{C}_n$? Of course, the extension of the subgroup metric, as a multiple of the cardinal of the symmetric difference of the sets of contacts, is straightforward. But, is it possible to generalize the subgroup representation to arbitrary contact structures? Notice that this representation, used *mutatis mutandis*, does not work: the extension of the mapping G to $\mathcal{C}_n$ is no longer injective, as the following result, whose easy proof we leave to the reader, shows.

Proposition 5 *For every* $\Gamma_1, \Gamma_2 \in \mathcal{C}_n$, $G(\Gamma_1) = G(\Gamma_2)$ *if and only if for every* $i \cdot j \in \Gamma_1$ *there exists a chain of contacts* $k_1 \cdot k_2, k_2 \cdot k_3, \ldots, k_{m-1} \cdot k_m$ *in* Γ_2 *with* $k_1 = i$ *and* $k_m = j$, *and vice versa, for every* $i \cdot j \in \Gamma_2$ *there is a similar chain of contacts in* Γ_1 *going from* i *to* j.

So, for instance, $\Gamma_1 = ([5], \{1{\cdot}3, 3{\cdot}5\})$ and $\Gamma_2 = ([5], \{1{\cdot}5, 3{\cdot}5\})$ have the same subgroup representation. Therefore, the extension of the subgroup metric to $\mathcal{C}_n$ is no longer a metric: $d_{sgr}(\Gamma_1, \Gamma_2) = 0$ does not imply $\Gamma_1 = \Gamma_2$.

On the other hand, Reidys and Stadler's subgroup representation of contact structures with unique bonds is framed in a general context, provided by the following result.

Proposition 6 **[17, Thm. 5]** *For every finite group* G, *the mapping*

$$\begin{array}{rccl} D: & \mathrm{Sub}(G) \times \mathrm{Sub}(G) & \to & \mathbb{R}^+ \\ & (H, K) & \mapsto & \ln |(K \cdot H)/(K \cap H)| \end{array}$$

is a metric on the set $\mathrm{Sub}(G)$ *of its subgroups.*

Therefore, any embedding of $\mathcal{C}_n$ into the set of subgroups of a finite group will induce a metric on it from this metric D. In §4.2 we shall use a similar approach to generalize the subgroup representation and metric on $\mathcal{U}_n$ to representations and metrics on $\mathcal{C}_n$, using polynomial ideals instead of permutation subgroups. Anyway, since they are surely not the only possible generalizations of the embedding $G : \mathcal{U}_n \to \mathrm{Sub}(S_n)$, we still want to propose the following problem:

Open problem 3 *To find other meaningful generalizations of the subgroup representation and metric to* $\mathcal{C}_n$.

3.3 Matrix representation and metric

Reidys and Stadler also used a representation of an RNA secondary structure as a symmetric and involutive matrix, previously introduced by Magarshak and his coworkers following a well-established tradition in spectral graph theory [5, p. 2], to propose a general method to compare RNA secondary structures. In this subsection we study an specific metric obtained through this method.

Magarshak *et al* [11, 14] represented a contact structure with unique bonds $\Gamma = ([n], Q)$ as an $n \times n$ complex symmetric matrix $S_\Gamma = (s_{i,j})_{i,j=1,\dots,n}$, with

$$s_{i,j} = \begin{cases} -1 & \text{if } i \neq j \text{ and } i \cdot j \in Q \\ 1 & \text{if } i = j \text{ and } i \cdot l \notin Q \text{ for every } l \\ 0 & \text{otherwise} \end{cases}$$

It is clear that this matrix S_Γ encodes the involution $\pi(\Gamma)$.

The main feature of these matrices that interested those authors was that they capture the compatibility of RNA sequences with contact structures. More specifically, if we represent the primary structure $\underline{b} = b_1 \dots b_n$ of an RNA molecule of length n as a vector

$$\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{C}^n$$

by encoding its bases by means of the rule

$$A \mapsto i,\ C \mapsto -1,\ G \mapsto 1,\ U \mapsto -i,$$

then the following key property is satisfied: for every RNA molecule $\underline{b} = b_1 b_2 \dots b_n$ of length n and for every $\Gamma = ([n], Q) \in \mathcal{U}_n$, if $\underline{x} \in \mathbb{C}^n$ is the vector representing $\underline{b}$, then $S_\Gamma \circ \underline{x} = \underline{x}$ if and only if $\underline{b}$ is *Watson-Crick compatible* with Γ, in the sense that if $j \cdot k \in Q$, then either $\{b_j, b_k\} = \{A, U\}$ or $\{b_j, b_k\} = \{C, G\}$.

Returning again to the matrix representation S_Γ of contact structures with unique bonds Γ, it is easy to check that $S_\Gamma^{-1} = S_\Gamma$ for every $\Gamma \in \mathcal{U}_n$. This implies that if we define, for every $\Gamma_1, \Gamma_2 \in \mathcal{U}_n$, the *transfer matrix* $T_{\Gamma_1,\Gamma_2} = S_{\Gamma_2} \circ S_{\Gamma_1}$, then these matrices satisfy, among other properties, that $T_{\Gamma_2,\Gamma_1} \circ S_{\Gamma_1} = S_{\Gamma_2}$, which motivates the adjective 'transfer,' and

$$T_{\Gamma_1,\Gamma_1} = \mathrm{Id}, \quad T_{\Gamma_2,\Gamma_1} = (T_{\Gamma_1,\Gamma_2})^{-1}, \quad T_{\Gamma_1,\Gamma_3} = T_{\Gamma_2,\Gamma_3} \circ T_{\Gamma_1,\Gamma_2}.$$

These properties led Reidys and Stadler to propose in [17, §3.2] the definition of metrics on $\mathcal{U}_n$ through

$$(\Gamma_1, \Gamma_2) \mapsto \|T_{\Gamma_1,\Gamma_2}\|,$$

where $\|\cdot\|$ stands for any *length function*[1] on the group $GL(n, \mathbb{C})$ of $n \times n$ invertible complex matrices. Taking in particular $\|A\| = \mathrm{rank}(A - \mathrm{Id})$, we obtain the *matrix metric* on $\mathcal{U}_n$

$$d_{mag}(\Gamma_1, \Gamma_2) = \mathrm{rank}(T_{\Gamma_1,\Gamma_2} - \mathrm{Id}).$$

[1] Actually, they proposed to use a matrix norm $\|\cdot\|$, but it is probably a misprint.

This metric turns out to be equal to the involution metric d_{inv} defined above.

Proposition 7 [18, Prop. 5] *For every* $\Gamma_1, \Gamma_2 \in \mathcal{U}_n$,

$$d_{mag}(\Gamma_1, \Gamma_2) = d_{inv}(\Gamma_1, \Gamma_2).$$

Open problem 4 *To investigate other metrics defined on* $\mathcal{U}_n$ *through other length functions on* $GL(n, \mathbb{C})$.

Now, although most of the hydrogen bonds in an RNA molecule form between between A and U and between C and G, a significant amount of bonds also form between other pairs of bases. This made Reidys and Stadler ask for a generalization of Magarshak's matrix representation of contact structures with unique bonds to allow a more general definition of compatibility between an RNA sequence and a contact structure. Our group has recently solved this problem [3], and our matrix representation of contact structures with unique bonds allows non-Watson-Crick base pairings as well as extended sets of ribonucleotides. Without entering into details on the codification of the bases, we represented a contact structure $\Gamma = ([n], Q)$ as a matrix $S_\Gamma = (s_{i,j})_{i,j=1,\ldots,n} \in GL(n, \mathbb{F}_{2^m})$ (where $\mathbb{F}_{2^m}$ stands for the finite field with 2^m elements, for any $m \geqslant 2$), with

$$s_{i,j} = \begin{cases} \alpha + 1 & \text{if } i \neq j \text{ and } i \cdot j \in Q \\ 0 & \text{if } i \neq j \text{ and } i \cdot j \notin Q \\ \alpha & \text{if } i = j \text{ and } i \cdot l \in Q \text{ for some } l \\ 1 & \text{if } i = j \text{ and } i \cdot l \notin Q \text{ for every } l \end{cases}$$

α being any generator of the cyclic multiplicative group $\mathbb{F}_{2^m} - \{0\}$. These matrices also satisfy that $S_\Gamma^{-1} = S_\Gamma$. If we still define transfer matrices by means of $T_{\Gamma_1,\Gamma_2} = S_{\Gamma_2} \circ S_{\Gamma_1}$ and we set $d_{mag}(\Gamma_1, \Gamma_2) = \operatorname{rank}(T_{\Gamma_1,\Gamma_2} - \mathrm{Id})$, then we have the following result.

Proposition 8 [3, Thms. 16, 17] *For every* $\Gamma_1, \Gamma_2 \in \mathcal{U}_n$*:*

(i) The product of the elements of the main diagonal of T_{Γ_1,Γ_2} *is* $\alpha^{2|Q_1 \triangle Q_2|}$.

(ii) $d_{mag}(\Gamma_1, \Gamma_2) = d_{inv}(\Gamma_1, \Gamma_2)$.

So, in this way we obtain again the involution metric (as well as, in some involved sense, the subgroup metric, provided we take m large enough). This motivates us to propose the following problem:

Open problem 5 *To generalize Magarshak's matrix representation of contact structures with unique bonds to arbitrary contact structures. Is it possible to obtain a generalization that allows to define a metric that extends* d_{mag}*, and hence* d_{inv}*, to* $\mathcal{C}_n$*?*

In our view, this approach provides a feasible way to generalize the involution representation (this would be given by the new matrix S_Γ) and metric to arbitrary contact structures.

4 Ideal-based representations and metrics

4.1 The general framework

In the previous section, we have surveyed Reidys and Stadler's metrics for contact structures with unique bonds based on the representation of these contact structures as elements or subgroups of a group, and we have discussed the possibility of extending them to arbitrary contact structures. In this subsection we present a general method to define metrics on $\mathcal{C}_n$, based on the representation of contact structures as monomial ideals and inspired in Reidys and Stadler's subgroup metric. In the next subsections we shall introduce several specific metrics on contact structures obtained in this way.

Let n be from now on an integer greater than 2. Let $\mathcal{M}(x_1, \ldots, x_n)$, or simply $\mathcal{M}(\underline{x})$, denote the set of all monomials in the variables $x_1, \ldots, x_n$. As usual, we shall denote a monomial $x_1^{\alpha_1} \cdots x_n^{\alpha_n} \in \mathcal{M}(\underline{x})$ by $x^{\underline{\alpha}}$, where $\underline{\alpha}$ stands for the n-tuple $(\alpha_1, \ldots, \alpha_n)$. The *total degree* of such a monomial $x^{\underline{\alpha}}$ is the sum $\sum_{i=1}^{n} \alpha_i$ of the variables' exponents in it. For every $m \geqslant 0$, let $\mathcal{M}(\underline{x})^{(m)}$ and $\mathcal{M}(\underline{x})_m$ be the sets of all monomials in $\mathcal{M}(\underline{x})$ of total degree m and of total degree $\leqslant m$, respectively. Recall that

$$|\mathcal{M}(\underline{x})^{(m)}| = \binom{n+m-1}{n-1} \text{ and } |\mathcal{M}(\underline{x})_m| = \binom{n+m}{n}.$$

Let $\mathbb{F}_2$ be the binary field $\mathbb{Z}/2\mathbb{Z}$ and $\mathbb{F}_2[x_1, \ldots, x_n]$, or simply $\mathbb{F}_2[\underline{x}]$, the ring of polynomials in the variables $x_1, \ldots, x_n$ with coefficients in $\mathbb{F}_2$. An ideal I of $\mathbb{F}_2[\underline{x}]$ is *monomial* when it is generated by a set of monomials. Let $\mathrm{Mon}(\mathbb{F}_2[\underline{x}])$ denote the set of monomial ideals of $\mathbb{F}_2[\underline{x}]$.

For every $I \in \mathrm{Mon}(\mathbb{F}_2[\underline{x}])$ and for every $m \geqslant 0$, let

$$M(I) = I \cap \mathcal{M}(\underline{x}), \quad M(I)^{(m)} = I \cap \mathcal{M}(\underline{x})^{(m)}, \quad M(I)_m = I \cap \mathcal{M}(\underline{x})_m.$$

Recall [6, §2.4] that, given a monomial ideal I generated by a set of monomials M, the monomials in $M(I)$ are exactly those that are divisible by some monomial in M and the polynomials in I are exactly the linear combinations (with coefficients in $\mathbb{F}_2$) of monomials in $M(I)$. Hilbert's basis theorem implies that every monomial ideal I has a unique non-redundant[2] finite set of monomials that generate it [6, Ex. 2.4.8]: we shall call it its *minimal basis*.

Given a monomial ideal $I \in \mathrm{Mon}(\mathbb{F}_2[\underline{x}])$, we shall denote by $\deg(I)$ the maximum of the total degrees of monomials in its minimal basis; in particular, $\deg(\{0\}) = \deg(\mathbb{F}_2[\underline{x}]) = 0$. Let

$$\mathrm{Mon}_m(\mathbb{F}_2[\underline{x}]) = \{I \in \mathrm{Mon}(\mathbb{F}_2[\underline{x}]) \mid \deg(I) \leqslant m\}.$$

Notice that if $k \leqslant l$, then $\mathrm{Mon}_k(\mathbb{F}_2[\underline{x}]) \subseteq \mathrm{Mon}_l(\mathbb{F}_2[\underline{x}])$.

In order to define metrics on $\mathcal{C}_n$ reminiscent of Reidys and Stadler's subgroup metric, let us introduce a family of mappings

$$D_m : \mathrm{Mon}(\mathbb{F}_2[\underline{x}]) \times \mathrm{Mon}(\mathbb{F}_2[\underline{x}]) \to \mathbb{R}, \qquad m \geqslant 1.$$

[2] In the sense that no element of it divides any other element of it.

defined in the following way. For every $n \geqslant 3$ and $m \geqslant 1$, let

$$\pi_m : \mathbb{F}[\underline{x}] \to \mathbb{F}_2[\underline{x}]/\langle \mathcal{M}(\underline{x})^{(m)} \rangle,$$

be the quotient ring homomorphism modulo the ideal $\langle \mathcal{M}(\underline{x})^{(m)} \rangle$. For every $I \in \mathrm{Mon}(\mathbb{F}[\underline{x}])$, let

$$\pi_m(I) = I/\langle \mathcal{M}(\underline{x})^{(m)} \rangle = (I + \langle \mathcal{M}(\underline{x})^{(m)} \rangle)/\langle \mathcal{M}(\underline{x})^{(m)} \rangle$$

be its image under π_m. Now, for every $m \geqslant 1$, and for every $I, J \in \mathrm{Mon}(\mathbb{F}_2[\underline{x}])$, let

$$D_m(I, J) = \log_2 \left| \frac{\pi_m(I) + \pi_m(J)}{\pi_m(I) \cap \pi_m(I)} \right|.$$

We have defined D_m in this way just to emphasize its connection, up to a constant factor, with the mapping D that underlies the subgroup metric introduced in Proposition 6. But, as it happens with the subgroup metric, this D_m has a simpler expression, which will be the one we shall use in proofs and computations. Indeed, a proof similar to that of Prop. 5 in [12] proves the following result.

Proposition 9 *For every $m \geqslant 1$ and for every $I, J \in \mathrm{Mon}(\mathbb{F}_2[\underline{x}])$,*

$$D_m(I, J) = |M(I)_{m-1} \Delta M(J)_{m-1}|.$$

Now, each D_m defines a metric on a certain subset of $\mathrm{Mon}(\mathbb{F}_2[\underline{x}])$.

Proposition 10 *For every $m \geqslant 1$, D_m is a metric on every $\mathrm{Mon}_k(\mathbb{F}_2[\underline{x}])$ with $k < m$.*

Proof. If $I \in \mathrm{Mon}(\mathbb{F}_2[\underline{x}])$ is generated by a set of monomials of total degree lower or equal than $m-1$, then this set of generators is contained in $M(I)_{m-1}$ and therefore $I = \langle M(I)_{m-1} \rangle$. This implies that the mapping (where $\mathcal{P}(\mathcal{M}(\underline{x})_{m-1})$ stands for the powerset of $\mathcal{M}(\underline{x})_{m-1}$)

$$\begin{array}{rcl} \mathrm{Mon}_{m-1}(\mathbb{F}_2[\underline{x}]) & \to & \mathcal{P}(\mathcal{M}(\underline{x})_{m-1}) \\ I & \mapsto & M(I)_{m-1} \end{array}$$

is an embedding. Then, since the cardinal of the symmetric difference is a metric on $\mathcal{P}(\mathcal{M}(\underline{x})_{m-1})$, it induces a metric on $\mathrm{Mon}_{m-1}(\mathbb{F}_2[\underline{x}])$ through this embedding, and hence on every $\mathrm{Mon}_k(\mathbb{F}_2[\underline{x}])$ with $k \leqslant m-1$. ∎

The next proposition provides an alternative description of D_m. It is easily obtained from Proposition 9 (cf. [12, Cor. 6]) and it is useful in the practical computation of these mappings D_m. In it, and for every ideal $I \in \mathrm{Mon}(\mathbb{F}[\underline{x}])$, $H_I : \mathbb{N} \to \mathbb{N}$ denotes the mapping defined by

$$H_I(m) = |\mathcal{M}(\underline{x})_m - M(I)_m|, \quad m \in \mathbb{N};$$

i.e., $H_I(m)$ is the number of monomials of total degree $\leqslant m$ that do not belong to I. This mapping is called the (affine) *Hilbert function* of I and it can be computed explicitly from any given finite set of monomial generators of I.

Proposition 11 *For every $m \geqslant 1$ and for every $I, J \in \mathrm{Mon}(\mathbb{F}_2[\underline{x}])$,*

$$D_m(I,J) = H_I(m-1) + H_J(m-1) - 2H_{I+J}(m-1).$$

Several freely available computer algebra systems like, for instance, CoCoA [2] or Macaulay [9], compute Hilbert functions. Therefore, this description of $D_m(I,J)$ provides a simple way to compute its value using one of these systems.

A monomial ideal I is *radical* when all monomials in its minimal basis are square-free. For every $m \geqslant 1$, let $\mathrm{RMon}_m(\mathbb{F}_2[\underline{x}])$ denote the set of radical ideals in $\mathrm{Mon}_m(\mathbb{F}_2[\underline{x}]) - \{\mathbb{F}_2[\underline{x}]\}$, i.e., excluding the total ideal. For non-total radical monomial ideals we can use a 'smaller' metric.

Proposition 12 *For every $m \geqslant 2$ and for every $I, J \in \mathrm{Mon}(\mathbb{F}_2[\underline{x}])$, let*

$$\widehat{D}_m(I,J) = D_m(I,J) - D_{m-1}(I,J) = |M(I)^{(m-1)} \Delta M(J)^{(m-1)}|.$$

Then, $\widehat{D}_m$ is a metric on the set $\mathrm{RMon}_k(\mathbb{F}_2[\underline{x}])$ for every $k < m$.

Proof. Let I, J be two different ideals in $\mathrm{RMon}_{m-1}(\mathbb{F}_2[\underline{x}])$ and assume that I is not contained into J. Then, there exists an element $x^{\underline{\alpha}}$ of the minimal basis of I that does not belong to J. Let $x^{\underline{\beta}}$ be a monomial of total degree $m-1$ obtained from $x^{\underline{\alpha}}$ by increasing the exponent of one of the variables appearing explicitly in it by $m-1$ minus the total degree of $x^{\underline{\alpha}}$ (every such $x^{\underline{\alpha}}$ has some variable, because $I \neq \mathbb{F}_2[\underline{x}]$ by assumption). Then, $x^{\underline{\beta}} \in M(I)^{(m-1)}$, because it is a multiple of an element of the minimal basis of I, but $x^{\underline{\beta}} \notin M(J)^{(m-1)}$, because if it were divisible by some monomial in the minimal basis of J, all of them square-free, then $x^{\underline{\alpha}}$ would also be divisible by this element, against the assumption that $x^{\underline{\alpha}} \notin J$. This shows that $M(I)^{(m-1)} \neq M(J)^{(m-1)}$. Therefore, the mapping

$$\begin{array}{rcl} \mathrm{RMon}_{m-1}(\mathbb{F}_2[\underline{x}]) & \to & \mathcal{P}(\mathcal{M}(\underline{x})^{(m-1)}) \\ I & \mapsto & M(I)^{(m-1)} \end{array}$$

is an embedding and hence the cardinal of the symmetric difference of subsets of $\mathcal{M}(\underline{x})^{(m-1)}$ induces, through this embedding, a metric on every $\mathrm{RMon}_k(\mathbb{F}_2[\underline{x}])$ with $k \leqslant m-1$. ∎

Notice that the mapping $I \mapsto M(I)^{(m-1)}$ does not define an embedding of the whole $\mathrm{Mon}_{m-1}(\mathbb{F}_2[\underline{x}]) - \{\mathbb{F}_2[\underline{x}]\}$ in $\mathcal{P}(\mathcal{M}(\underline{x})^{(m-1)})$. For instance, if we let $I = \langle x_1 \rangle$ and $J = \langle x_1^2, x_1x_2, \ldots, x_1x_n \rangle$, then $M(I)^{(m)} = M(J)^{(m)}$ for every $m \geqslant 2$. It is also clear that we must remove the total ideal if we want to get an embedding: $M(\langle 1 \rangle)^{(m)} = M(\langle x_1, \ldots, x_n \rangle)^{(m)}$ for every $m \geqslant 1$.

If necessary, from Proposition 11 one can obtain an expression for $\widehat{D}_m$ in terms of Hilbert functions that allows to compute it using a computer algebra system.

Now, let us return to contact structures. Propositions 10 and 12 provide two general methods to define metrics on $\mathcal{C}_n$. If we define some injective representation of contact structures as monomial ideals I in $\mathbb{F}_2[\underline{x}]$ with $\deg(I)$ bounded,

then we can use the metrics D_m, for sufficiently large m, to induce a family of metrics on $\mathcal{C}_n$. And if the ideals I associated to contact structures are radical, as it will always be the case in this paper, then we can also use the metrics $\widehat{D}_m$ with the same purpose. In the next subsections we explain several metrics on contact structures obtained in this way.

4.2 Edge ideal representation and metrics

In [12] we used the first approach explained in the last subsection to define a metric on $\mathcal{C}_n$ using the representation of contact structures as edge ideals. Edge ideals are quite a popular tool in commutative algebra to represent graphs and to study their properties [22] and by using them we obtained what have been, to our knowledge, the first metrics defined on arbitrary contact structures of a fixed length that are independent of any notion of graph edition. We recall these metrics in this subsection.

Let us start by defining the edge ideal of a contact structure.

Definition 2 *For every $\Gamma = ([n], Q) \in \mathcal{C}_n$, the* edge ideal I_Γ *of Γ is the ideal of $\mathbb{F}_2[\underline{x}]$ generated by the products of pairs of variables whose indices form a contact in Γ:*

$$I_\Gamma = \langle \{x_i x_j \mid i \cdot j \in Q\} \rangle.$$

It is clear that $\{x_i x_j \mid i \cdot j \in Q\}$ is a minimal basis of I_Γ consisting of square-free monomials, and therefore $I_\Gamma \in \mathrm{RMon}_2(\mathbb{F}_2[\underline{x}])$ (notice that $\deg(I_\Gamma)$ can be 0: if Γ is the empty secondary structure, in which case $I_\Gamma = \{0\}$). On the other hand, it is also clear that

$$\begin{array}{rcl} \mathcal{C}_n & \to & \mathrm{RMon}_2(\mathbb{F}_2[\underline{x}]) \\ \Gamma & \mapsto & I_\Gamma \end{array}$$

is an embedding: since the monomials in I_Γ are exactly those divisible by some $x_i x_j$ with $i \cdot j \in Q$, we have that

$$M(I_\Gamma)_2 = \{x_i x_j \mid i \cdot j \in Q\},$$

and hence Γ is uniquely determined by $M(I_\Gamma)_2$. Therefore, as a simple application of Propositions 9 to 11 (and using that $I_{\Gamma_1} + I_{\Gamma_2} = I_{\Gamma_1 \cup \Gamma_2}$) we obtain the following result.

Proposition 13 [12, §3] *For every $m \geqslant 3$, the mapping $d'_m : \mathcal{C}_n \times \mathcal{C}_n \to \mathbb{R}$ defined by*

$$d'_m(\Gamma_1, \Gamma_2) = D_m(I_{\Gamma_1}, I_{\Gamma_2})$$

is a metric. Moreover,

$$\begin{array}{rl} d'_m(\Gamma_1, \Gamma_2) & = |M(I_{\Gamma_1})_{m-1} \Delta M(I_{\Gamma_2})_{m-1}| \\ & = H_{I_{\Gamma_1}}(m-1) + H_{I_{\Gamma_2}}(m-1) - 2H_{I_{\Gamma_1 \cup \Gamma_2}}(m-1). \end{array}$$

Actually, we shall not use these metrics d'_m, but scalar factors of them. Let Γ_0 be the empty contact structure of length n and let Γ_1 be an RNA secondary structure of length n with only one contact, say $i \cdot j$. Then $I_{\Gamma_0} = \{0\}$ and $I_{\Gamma_1} = \langle x_i x_j \rangle$, and therefore, for every $m \geqslant 3$,

$$\begin{aligned} d'_m(\Gamma_0, \Gamma_1) &= |M(I_{\Gamma_0})_{m-1} \Delta M(I_{\Gamma_1})_{m-1}| = |M(\langle x_i x_j \rangle)_{m-1}| \\ &= |\{x_i x_j \cdot x^{\underline{\alpha}} \mid x^{\underline{\alpha}} \in \mathcal{M}(\underline{x})_{m-3}\}| = |\mathcal{M}(\underline{x})_{m-3}| = \tbinom{n+m-3}{n}. \end{aligned}$$

Taking 1 as the 'natural' value for the distance between Γ_0 and Γ_1, we divide every metric d'_m on $\mathcal{C}_n$ by $\binom{n+m-3}{n}$.

Definition 3 *For every $m \geqslant 3$, the* edge ideal mth metric *on $\mathcal{C}_n$ is*

$$d_m(\Gamma_1, \Gamma_2) = \frac{1}{\binom{n+m-3}{n}} d'_m(\Gamma_1, \Gamma_2), \quad \Gamma_1, \Gamma_2 \in \mathcal{C}_n.$$

In [12] we have explicitly computed the metric d_m on $\mathcal{U}_n$ and $\mathcal{C}_n$ for several low indices m and we have discussed several examples that show their possible range of application. Let us simply recall here some of the formulas we obtained there.

Proposition 14 *For every $\Gamma_1, \Gamma_2 \in \mathcal{C}_n$, $d_3(\Gamma_1, \Gamma_2) = |Q_1 \Delta Q_2|$.*

Thus, d_3 generalizes the subgroup metric to arbitrary contact structures. Actually, this generalization is deeper than its rough value, since it can be easily seen that, for every $\Gamma \in \mathcal{U}_n$, $\pi_3(I_\Gamma)$ (as a group with the sum of congruence classes) is isomorphic to $G(\Gamma)$.

For every $\Gamma_1, \Gamma_2 \in \mathcal{U}_n$, let Ω_m denote the number of open orbits of $\Gamma_1 \Delta \Gamma_2$ with at least m nodes, and $\Theta^{(m)}$ the number of closed orbits of $\Gamma_1 \Delta \Gamma_2$ with exactly m nodes.

Proposition 15 *For every $\Gamma_1, \Gamma_2 \in \mathcal{U}_n$,*

$$d_4(\Gamma_1, \Gamma_2) = |Q_1 \Delta Q_2| - \frac{2}{n+1}(|Q_1 \Delta Q_2| - \Omega_2).$$

Therefore, on $\mathcal{U}_n$, the metric d_4 increases with the cardinal of $Q_1 \Delta Q_2$, but decreases with the number of *angles* in $\Gamma_1 \Delta \Gamma_2$, i.e., of pairs of contacts in $Q_1 \Delta Q_2$ that share a node. Indeed, it is easy to prove that this number of angles is exactly $|Q_1 \Delta Q_2| - \Omega_2$.

Proposition 16 *For every $\Gamma_1, \Gamma_2 \in \mathcal{U}_n$,*

$$\begin{aligned} d_5(\Gamma_1, \Gamma_2) = \; & |Q_1 \Delta Q_2| - \tfrac{1}{\binom{n+2}{2}} \Big(2\tbinom{|Q_1 \cup Q_2|}{2} - \tbinom{|Q_1|}{2} - \tbinom{|Q_2|}{2} \\ & + 2(n-1)(|Q_1 \Delta Q_2| - \Omega_2) + 2(\Omega_3 + \Theta^{(4)}) \Big) \end{aligned}$$

In this case, the term $2\binom{|Q_1\cup Q_2|}{2}-\binom{|Q_1|}{2}-\binom{|Q_2|}{2}$ makes the value of $d_5(\Gamma_1,\Gamma_2)$ depend not only on the cardinal and structure of the set $Q_1\Delta Q_2$, but also on $|Q_1\cap Q_2|$. More specifically, it turns out that if $\Gamma_1,\Gamma_2,\Gamma_1',\Gamma_2'\in\mathcal{U}_n$ are such that $Q_1-Q_2=Q_1'-Q_2'$, $Q_2-Q_1=Q_2'-Q_1'$ and $|Q_1\cap Q_2|<|Q_1'\cap Q_2'|$, then $d_5(\Gamma_1,\Gamma_2)>d_5(\Gamma_1',\Gamma_2')$; i.e., the greater the set of contacts they share, the closer they are.

To end this part of this subsection, let us mention that, for every $m\leqslant n$, the value of the edge ideal mth metric involves a parameter Ω_m or $\Theta^{(m)}$ that is not involved in the value of edge ideal metrics with lower indices, and therefore each such metric captures a different notion of similarity.

Now, one could also use the edge ideal representation of arbitrary contact structures to define metrics on $\mathcal{C}_n$ using the metrics $\widehat{D}_m$, $m\geqslant 3$, on $\mathrm{RMon}_2(\mathbb{F}_2[\underline{x}])$. In this case, we would have that, for every $m\geqslant 3$, the mapping defined by

$$\widehat{d}'_m(\Gamma_1,\Gamma_2)=|M(I_{\Gamma_1})^{(m-1)}\Delta M(I_{\Gamma_2})^{(m-1)}|=d'_m(\Gamma_1,\Gamma_2)-d'_{m-1}(\Gamma_1,\Gamma_2),$$

is a metric on $\mathcal{C}_n$. If Γ_0 is the empty secondary structure of length n and Γ_1 is a secondary structure of length n with only one contact, then

$$\widehat{d}'_m(\Gamma_0,\Gamma_1)=\binom{n+m-3}{n}-\binom{n+m-4}{n}=\binom{n+m-4}{n-1}.$$

Therefore, after normalizing $\widehat{d}'_m$ by this factor, we can define, for every $m\geqslant 3$, a *second edge ideal mth metric* on $\mathcal{C}_n$ by

$$\widehat{d}_m(\Gamma_1,\Gamma_2)=\frac{1}{\binom{n+m-4}{n-1}}\widehat{d}'_m(\Gamma_1,\Gamma_2),\quad \Gamma_1,\Gamma_2\in\mathcal{C}_n.$$

It is straightforward to check that, for every $\Gamma_1,\Gamma_2\in\mathcal{C}_n$,

$$\widehat{d}_3(\Gamma_1,\Gamma_2)=d_3(\Gamma_1,\Gamma_2)=|Q_1\Delta Q_2|$$

and hence, if you want, the family of metrics $(\widehat{d}_m)_{m\geqslant 3}$ also generalizes the subgroup metric. We also have that, for every $\Gamma_1,\Gamma_2\in\mathcal{U}_n$,

$$\widehat{d}_4(\Gamma_1,\Gamma_2)=|Q_1\Delta Q_2|-\frac{2}{n}(|Q_1\Delta Q_2|-\Omega_2).$$

And so on.

It is clear that the family $(\widehat{d}_m)_{m\geqslant 3}$ captures (globally) the same information as the family $(d_m)_{m\geqslant 3}$, but we consider it may deserve a deep study. For instance, we might ask ourselves whether every $\widehat{d}_m$ captures the same information as the corresponding d_m, or what is $\lim_{m\to\infty}\widehat{d}_m(\Gamma_1,\Gamma_2)$.

Open problem 6 *To study the metrics $(\widehat{d}_m)_{m\geqslant 3}$ on $\mathcal{C}_n$.*

4.3 Clique ideal representation and metrics

Edge ideals are not the unique possible monomial ideal representations of graphs, and hence of arbitrary contact structures. Other ideals, like path and toric ideals are studied in [22] and could be used with this purpose. In this subsection we discuss another possible such representation, which, again to our knowledge, is new in the literature.

Recall that a *maximal clique* of a graph is a maximal complete subgraph of it. Such a clique is *non-trivial* when it has at least two nodes. For every $\Gamma = ([n], Q) \in \mathcal{C}_n$, let $\mathrm{Cli}(\Gamma)$ be its set of non-trivial maximal cliques. Notice that the knowledge of $\mathrm{Cli}(\Gamma)$ completely determines Γ: $i{\cdot}j \in Q$ if and only if i, j belong simultaneously to some non-trivial maximal clique.

Definition 4 *For every $\Gamma = ([n], Q) \in \mathcal{C}_n$, the* clique ideal K_Γ *of Γ is the ideal of $\mathbb{F}_2[\underline{x}]$ generated by the set of monomials consisting of one square-free monomial $x_{i_1} \cdots x_{i_k}$ for each non-trivial maximal clique $\{i_1, \ldots, i_k\}$ of Γ:*

$$I_\Gamma = \left\langle \left\{ x_{i_1} \cdots x_{i_k} \mid \{i_1, \ldots, i_k\} \in \mathrm{Cli}(\Gamma) \right\} \right\rangle.$$

As in the case of edge ideals, the definition of maximal clique entails that the set of monomials generating K_Γ specified in this definition is a minimal basis of K_Γ, and hence $\deg(K_\Gamma) \leqslant \lceil n/2 \rceil$: the number $n/2$ is due to the fact that a contact structure cannot contain contacts between consecutive bases, and the upper bound $\lceil n/2 \rceil$ is reached for instance when all odd-numbered nodes contact with each other forming a clique. On the other hand, it should be noticed that if $\Gamma \in \mathcal{U}_n$, then $\mathrm{Cli}(\Gamma)$ consists of those 2-element sets $\{i, j\}$ such that $i{\cdot}j \in Q$ and therefore in this case $K_\Gamma = I_\Gamma$.

Now, it is again clear that

$$\begin{array}{rcl} \mathcal{C}_n & \to & \mathrm{RMon}_{\lceil n/2 \rceil}(\mathbb{F}_2[\underline{x}]) \\ \Gamma & \mapsto & K_\Gamma \end{array}$$

is an embedding. Therefore, a simple application of Propositions 9 and 10 yields the following result.

Proposition 17 *For every $n \geqslant 3$ and for every $m \geqslant \lceil n/2 \rceil + 1$, the mapping $d^{c\prime}{}_m : \mathcal{C}_n \times \mathcal{C}_n \to \mathbb{R}$ defined by*

$$d^{c\prime}{}_m(\Gamma_1, \Gamma_2) = D_m(K_{\Gamma_1}, K_{\Gamma_2}) = |M(K_{\Gamma_1})_{m-1} \Delta M(K_{\Gamma_2})_{m-1}|$$

is a metric.

Moreover, Proposition 11 provides an expression for $d^{c\prime}{}_m$ in terms of Hilbert functions, which in this case simply says

$$d^{c\prime}{}_m(\Gamma_1, \Gamma_2) = H_{K_{\Gamma_1}}(m-1) + H_{K_{\Gamma_2}}(m-1) - 2H_{K_{\Gamma_1}+K_{\Gamma_2}}(m-1)$$

because, for clique ideals, it is in general false that $K_{\Gamma_1 \cup \Gamma_2} = K_{\Gamma_1} + K_{\Gamma_2}$.

As in the edge ideal metric case, we must normalize the metric ${d^c}'_m$ in order to keep the figures acceptably small. Since if $\Gamma_1, \Gamma_2 \in \mathcal{U}_n$, then $K_{\Gamma_1} = I_{\Gamma_1}$ and $K_{\Gamma_2} = I_{\Gamma_2}$, and hence in this case

$$d'_m(\Gamma_1, \Gamma_2) = {d^c}'_m(\Gamma_1, \Gamma_2),$$

it is clear that we must divide ${d^c}'_m$ on $\mathcal{C}_n$ again by $\binom{n+m-3}{n}$ if we want the distance from an empty contact structure to a contact structure of the same length and a single contact to be 1.

Definition 5 *For every $m \geqslant \lceil n/2 \rceil + 1$, the* clique ideal mth metric *on $\mathcal{C}_n$ is*

$$d^c_m(\Gamma_1, \Gamma_2) = \frac{1}{\binom{n+m-3}{n}} {d^c}'_m(\Gamma_1, \Gamma_2), \quad \Gamma_1, \Gamma_2 \in \mathcal{C}_n.$$

The value of a clique ideal metric depends on the sets of cliques, rather than on the sets of contacts. We shall show it with a simple example. Let $n = 5$, and consider the contact structures of length 5

$$\Gamma_1 = ([5], \{1{\cdot}3, 3{\cdot}5, 1{\cdot}5\}), \quad \Gamma_2 = ([5], \{1{\cdot}3, 3{\cdot}5\}), \quad \Gamma_3 = ([5], \{1{\cdot}3\}).$$

Thus, Γ_1 contains a clique $\{1, 3, 5\}$, Γ_2 is obtained from Γ_1 by removing a contact, which splits the clique into two, and Γ_3 is obtained from Γ_2 by removing a second contact, leaving only one contact that forms a clique by itself. We have that

$$\begin{aligned} I_{\Gamma_1} &= \langle x_1x_3, x_1x_5, x_3x_5 \rangle, & K_{\Gamma_1} &= \langle x_1x_3x_5 \rangle, \\ I_{\Gamma_2} = K_{\Gamma_2} &= \langle x_1x_3, x_1x_5 \rangle, & I_{\Gamma_3} = K_{\Gamma_3} &= \langle x_1x_3 \rangle. \end{aligned}$$

In this case $\lceil n/2 \rceil + 1 = 4$, and a simple computation yields the following figures:

$$\begin{aligned} d^c_4(\Gamma_1, \Gamma_2) &= 10/6, & d_4(\Gamma_1, \Gamma_2) &= 5/6, \\ d^c_4(\Gamma_1, \Gamma_3) &= 5/6, & d_4(\Gamma_1, \Gamma_3) &= 10/6 \\ d^c_4(\Gamma_2, \Gamma_3) &= 5/6, & d_4(\Gamma_2, \Gamma_3) &= 5/6 \end{aligned}$$

It is interesting to compare the behavior of d_4 and d^c_4 on these contact structures. Under d_4, Γ_2 is closer to Γ_1 than Γ_3, which corresponds to the closeness of the sets of contacts. But under d^c_4, Γ_3 is closer to Γ_1 than Γ_2, which seems to correspond to the closeness of the sets of cliques.

Since the minimal basis of K_Γ used in its definition consists of square-free monomials, one can also use the metrics $\widehat{D}_m$ with $m \geqslant \lceil n/2 \rceil + 1$ to compare these ideals and hence arbitrary contact structures. More specifically, for every $m \geqslant \lceil n/2 \rceil + 1$ we have a metric on $\mathcal{C}_n$ defined by

$${\widehat{d^c}}'_m(\Gamma_1, \Gamma_2) = \widehat{D}_m(K_{\Gamma_1}, K_{\Gamma_2})$$

and, after dividing it by $\binom{n+m-4}{n-1}$ to normalize it, we obtain a second family of clique ideal metrics $(\widehat{d^c_m})_{m \geqslant \lceil n/2 \rceil + 1}$ on $\mathcal{C}_n$.

We believe that these metrics d^c_m and $\widehat{d^c_m}$ can be useful in the comparison of protein contact structures, which are tightly packed and where the similarity of their cliques' structure can be more relevant than that of their contacts' structure. Thus, we propose the following open problem:

Open problem 7 *To develop and study these metrics based on clique ideal representations.*

In our view, explicit descriptions for the metrics d_m^c and $\widehat{d_m^c}$ similar in spirit to those given in [12] for edge ideal metrics, will be difficult to obtain, but they could be a key tool to understand these metrics.

To end this subsection, let us mention again that other representations of contact structures as monomial ideals, some of them already extensively studied in commutative algebra, could be used to define metrics on $\mathcal{C}_n$.

Open problem 8 *To define and study other metrics on $\mathcal{C}_n$ derived from the representation of contact structures as monomial ideals.*

4.4 Primary-edge ideal representation and metrics

All metrics introduced so far compare contact structures without paying attention to the monomers assigned to each node by the primary structure of an specific biopolymer. But, at a higher level of precision, it may be interesting to take into account not only the contact structure of a biomolecule, but also its primary structure, as most edit distances on RNA molecules of arbitrary length do [13]. Now, it is possible to introduce the primary structure of biomolecules in our ideal-based representations, at the price of increasing the number of variables in the polynomial ring we work in. We shall show in this subsection one possible way to do it, based on a generalization of edge ideals. For simplicity, we shall only consider here RNA contact structures, not necessarily with unique bonds, but it should be clear that this approach can be developed for structures of biomolecules whose primary structures are words on any fixed alphabet, from the binary $\{\mathrm{Pur}, \mathrm{Pyr}\}$ alphabet for RNA molecules to different alphabets to represent aminoacids or equivalence classes of them for proteins.

Definition 6 *An* RNA whole structure *of* length n *is a pair* $\Sigma = (\underline{b}, \Gamma)$ *where* $\underline{b} \in \{A, C, G, U\}^n$ *is a primary structure of an RNA molecule and* $\Gamma = ([n], Q)$ *is a contact structure, both of length* n*.*

We could impose compatibility restrictions between the primary and contact structure in RNA whole structures, as was discussed in §3.3, but we shall not take them into account here. Let $\mathcal{WRNA}_n$ denote the set of all RNA whole structures of length n.

We shall work now with a set of polynomial variables

$$a_1, \ldots, a_n, c_1, \ldots, c_n, g_1, \ldots, g_n, u_1, \ldots, u_n,$$

but, for simplicity, we shall still denote by $\mathcal{M}(\underline{x})$ and $\mathbb{F}_2[\underline{x}]$ the set of monomials and the ring of polynomials in these variables with coefficients in $\mathbb{F}_2$.

Definition 7 *For every* $\Sigma = (\underline{b}, \Gamma) \in \mathcal{WRNA}_n$*, the* primary-edge ideal J_Σ *of* Σ *is the ideal of* $\mathbb{F}_2[\underline{x}]$ *generated by the following monomials:*

(1) If $i \in [n]$ is isolated in Γ and $b_i = X \in \{A, C, G, U\}$, then the set of generators of J_Σ contains the corresponding variable $x_i \in \{a_i, c_i, g_i, u_i\}$.

(2) If $i \cdot j \in Q$ and $b_i = X$, $b_j = Y$, with $X, Y \in \{A, C, G, U\}$, then the set of generators of J_Σ contains the product $x_i y_j$ of the corresponding pair of variables $x_i, y_i \in \{a_i, c_i, g_i, u_i\}$.

It is clear that the set of generators of J_Σ given in this definition is a minimal basis of it, and that

$$\begin{array}{rcl} \mathcal{WRNA}_n & \to & \mathrm{RMon}_2(\mathbb{F}_2[\underline{x}]) \\ \Sigma & \mapsto & J_\Sigma \end{array}$$

is an embedding. Therefore, we can use again Propositions 9 and 10 to define a family of metrics on $\mathcal{WRNA}_n$.

Proposition 18 *For every $n \geqslant 3$ and for every $m \geqslant 3$, the mapping*

$$\tilde{d}_m : \mathcal{WRNA}_n \times \mathcal{WRNA}_n \to \mathbb{R}$$

defined by

$$\tilde{d'}_m(\Sigma_1, \Sigma_2) = D_m(J_{\Sigma_1}, J_{\Sigma_2}) = |M(J_{\Sigma_1})_{m-1} \Delta M(J_{\Sigma_2})_{m-1}|$$

is a metric.

As we already encountered in the previous two subsections, these metrics yield artificially large figures (even larger now, because of the increase in the number of variables), and therefore they must be normalized. But now there are several possible natural normalizations. Here, we shall assign 1 to the distance from an empty RNA whole structure to an empty RNA whole structure that differs from it in only one position of the primary structure.

So, let Γ_0 be the empty RNA secondary structure of length n, and let $\Sigma_1 = (\underline{b}_1, \Gamma_0), \Sigma_2 = (\underline{b}_2, \Gamma_0) \in \mathcal{WRNA}_n$ be such that $\underline{b}_1$ and $\underline{b}_2$ differ only in one position; without any loss of generality, we shall assume that $\underline{b}_1 = AA \ldots A$ and $\underline{b}_2 = CA \ldots A$. Then

$$J_{\Sigma_1} = \langle a_1, a_2, \ldots, a_n \rangle, \quad J_{\Sigma_2} = \langle c_1, a_2, \ldots, a_n \rangle.$$

In this case $M(J_{\Sigma_1}) \Delta M(J_{\Sigma_2})$ consists of those monomials that are either divisible by a_1 and not divisible by $c_1, a_2, \ldots, a_n$ or divisible by c_1 and not divisible by $a_1, a_2, \ldots, a_n$. Therefore

$$\begin{aligned} \tilde{d'}_m(\Sigma_1, \Sigma_2) &= |M(J_{\Sigma_1})_{m-1} \Delta M(J_{\Sigma_2})_{m-1}| \\ &= |\mathcal{M}(a_1, g_1, u_1, c_2, g_2, u_2, \ldots, c_n, g_n, u_n)_{m-1} \\ &\qquad -\mathcal{M}(g_1, u_1, c_2, g_2, u_2, \ldots, c_n, g_n, u_n)_{m-1}| \\ &\quad +|\mathcal{M}(c_1, g_1, u_1, c_2, g_2, u_2, \ldots, c_n, g_n, u_n)_{m-1} \\ &\qquad -\mathcal{M}(g_1, u_1, c_2, g_2, u_2, \ldots, c_n, g_n, u_n)_{m-1}| \\ &= 2\left(\tbinom{3n+m-1}{3n} - \tbinom{3n+m-2}{3n-1}\right) = 2\tbinom{3n+m-2}{3n}. \end{aligned}$$

Definition 8 *For every $n \geqslant 3$ and for every $m \geqslant 3$, the* primary-edge ideal mth metric *on $\mathcal{WRNA}_n$ is*

$$\tilde{d}_m(\Sigma_1, \Sigma_2) = \frac{1}{2\binom{3n+m-2}{3n}} \tilde{d}'_m(\Sigma_1, \Sigma_2), \quad \Sigma_1, \Sigma_2 \in \mathcal{WRNA}_n.$$

As an example, let us compute now the distance $\tilde{d}_m$ from an empty RNA whole structure to an RNA whole structure with the same primary structure and one contact. Without any loss of generality, let $\Sigma_1 = (\underline{b}_0, \Gamma_0)$ be as before, and let $\Sigma_3 = (\underline{b}_0, \Gamma_1)$ be the RNA whole structure with the same primary structure and $\Gamma_1 = \{1 \cdot n\}$. Then

$$J_{\Sigma_1} = \langle a_1, a_2, \ldots, a_{n-1}, a_n \rangle, \quad J_{\Sigma_3} = \langle a_1 a_n, a_2 \ldots, a_{n-1} \rangle.$$

Since $J_{\Sigma_3} \subseteq J_{\Sigma_1}$, in this case

$$M(J_{\Sigma_1}) \Delta M(J_{\Sigma_3}) = M(J_{\Sigma_1}) - M(J_{\Sigma_3})$$

consists of those monomials that are either divisible by a_1 and not divisible by $a_2, \ldots, a_n$ or divisible by a_n and not divisible by $a_1, a_2, \ldots, a_{n-1}$. Therefore

$$\begin{aligned}
\tilde{d}'_m(\Sigma_1, \Sigma_3) &= |M(J_{\Sigma_1})_{m-1} - M(J_{\Sigma_3})_{m-1}| \\
&= |\mathcal{M}(a_1, c_1, g_1, u_1, c_2, g_2, u_2, \ldots, c_n, g_n, u_n)_{m-1} \\
&\qquad -\mathcal{M}(g_1, c_1, u_1, c_2, g_2, u_2, \ldots, c_n, g_n, u_n)_{m-1}| \\
&\quad +|\mathcal{M}(c_1, g_1, u_1, c_2, g_2, u_2, \ldots, a_n, c_n, g_n, u_n)_{m-1} \\
&\qquad -\mathcal{M}(c_1, g_1, u_1, c_2, g_2, u_2, \ldots, c_n, g_n, u_n)_{m-1}| \\
&= 2\left(\tbinom{3n+m}{3n+1} - \tbinom{3n+m-1}{3n}\right) = 2\tbinom{3n+m-1}{3n+1}
\end{aligned}$$

and hence

$$\tilde{d}_m(\Sigma_1, \Sigma_3) = \frac{\binom{3n+m-1}{3n+1}}{\binom{3n+m-2}{3n}} = \frac{3n+m-1}{3n+1} = 1 + \frac{m-2}{3n+1} > 1$$

Notice that, as n grows relatively to m, this value converges to 1.

Of course, another family of primary-edge ideal metrics on $\mathcal{WRNA}_n$ can be defined using these primary-edge ideal representations of RNA whole structures and, now, the metrics $\widehat{D}_m$ introduced in Proposition 12. We leave to the interested reader to write up the definitions and to find the normalization factor.

Open problem 9 *To develop and study these metrics based on primary-edge ideal representations.*

Of course, primary-edge ideals are not the only possible way to represent simultaneously the primary and the contact structure of a biomolecule.

Open problem 10 *To introduce other representations of biomolecular whole structures and to use them to define new metrics.*

5 Conclusion

In this paper we have surveyed the state of the art of the algebraic representation of biomolecular structures and the use of these algebraic models in the definition of metrics on biomolecular contact structures, a field of research opened by Reidys and Stadler in 1996. We have taken the opportunity to recall several interesting open problems in this area. We are currently working on some of these problems, but we also would be glad to see them solved by anyone else.

We have focused on distances derived from algebraic models, skipping whole areas of research in the comparison of RNA structures like, for instance, edit distances or topological indices. Each one of these topics would deserve a survey by itself. And we have mainly considered representations of contact structures based on groups and on polynomial (actually, monomial) ideals. Of course, there are probably other models related to other algebraic structures waiting to be discovered and that would lead to new, interesting metrics.

Although we have focused on contact structures of the same length, whose comparison has several specific applications, it is clear that the comparison of contact structures of arbitrary length is more interesting. The problem of defining algebraic models of contact structures (or even of RNA secondary structures) of variable length, from which to derive abstract metrics for contact structures of possibly different lengths in a way independent of any notion of graph edition, remains as open as it was when Reidys and Stadler posed it in 1996.

And finally, a last, and possibly the most important, question remains open. What is the biochemical relevance of all these representations and metrics? At this point in time, no metric seems to be more suitable than any other one to compare biomolecular contact structures, even for specific types of structures and with specific purposes. A thorough analysis of all metrics introduced so far, both from the theoretical point of view and through extensive simulations, could show whether there is one metric that, for instance, captures more closely the probability of the fact that one structure admits one single-step mutation that transforms it into another structure. This is a field where mathematicians, computer scientists, and biologists seem predetermined to collaborate.

Acknowledgements. We would like to thank R. Alberich, J. Casasnovas, M. Llabrés, J. Miró and G. Valiente for many discussions and input along the last years on the topic of this paper. We want to thank specially M. Llabrés and J. Miró for their comments on a first version of this paper, which have led to a substantial improvement of it. We would like to thank the organizers of the Biomolecular Mathematics Special Session of the RSME-AMS First Joint Meeting (Sevilla, june 2003) for the opportunity to present some of the contents of this paper for the first time. This work has been partially supported by the Spanish DGES and the EU program FEDER, project BFM2003-00771 ALBIOM.

References

[1] R. T. Batey, R. P. Rambo, J. A. Doudna, Tertiary motifs and folding of RNA, *Angew. Chem. Int. Ed.* **38** (1999), 2326–2343.

[2] A. Capani, G. Niesi, L. Robbiano, CoCoA, a system for doing Computations in Commutative Algebra, available via anonymous ftp from `cocoa.dima.unige.it`.

[3] J. Casasnovas, J. Miró, F. Rosselló, On the algebraic representation of RNA secondary structures with GU pairs, *J. Math. Biol.* **47** (2003), 1–22.

[4] H. S. Chan, K. A. Dill, Sequence space soup of proteins and copolymers, *J. Chem. Phys.* **95** (1991), 3775–3779.

[5] F. R. K. Chung, *Spectral Graph Theory*, CBMS vol. 92, AMS (1994).

[6] D. Cox, J. Little, D. O'Shea, *Ideals, Varieties, and Algorithms*, Springer-Verlag (second ed., 1997).

[7] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yeo, P. D. Thomas, H. S. Chan, Principles of protein folding: A perspective from simple exact models, *Prot. Sci.* **4** (1995), 561-602.

[8] W. Fontana, P. Schuster, Shaping space: The possible and the attainable in RNA genotype-phenotype mapping, *J. Theor. Biol.* **194** (1998), 491–515.

[9] D. R. Grayson, M. Stillman, Macaulay 2, software system available at `http://www.math.uiuc.edu/Macaulay2/`.

[10] I. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, P. Schuster, Fast folding and comparison of RNA secondary structures, *Monatsh. Chem.* **125** (1994), 167–188.

[11] A. Kister, Y. Magarshak, J. Malinsky, The theoretical analysis of the process of RNA molecule self-assembly, *BioSystems* **30** (1993), 31–48.

[12] M. Llabrés, F. Rosselló. A new family of metrics for biopolymer contact structures, *Computational Biology and Chemistry* **28** (2004), 21–37.

[13] B. Ma, L. Wang, K. Zhang, Computing similarity between RNA structures, *Theor. Comp. Sc.* **276** (2002), 111-132.

[14] Y. Magarshak, C. J. Benham, An algebraic representation of RNA secondary structures, *J. of Biomol. Struct. & Dyn.* **10** (1992) 465–488.

[15] P. B. Moore, Structural motifs in RNA, *Annu. Rev. Biochem.* **68** (1999), 287–300.

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

[17] C. Reidys, P. F. Stadler, Bio-molecular shapes and algebraic structures, *Comp. & Chem.* **20** (1996), 85–94.

[18] F. Rosselló, On Reidys and Stadler's metrics for RNA secondary structures, to appear in *Math. and Comp. Mod.*

[19] P. Schuster, W. Fontana, P. Stadler, I. Hofacker, From sequences to shapes and back: A case study in RNA secondary structures, *Proc. Roy. Soc. B* **255** (1994), 279–284.

[20] P. Schuster, P. F. Stadler, Discrete models of biopolymers, to appear in *Handbook of Computational Chemistry* (M.J.C. Crabbe, M. Drew and A. Konopka, eds.), Marcel Dekker (in press); see also Univ. Wien TBI Preprint No. pks-99-012 (1999).

[21] B. Shapiro, K. Zhang, Comparing multiple RNA secondary structures using tree comparisons, *CABIOS* **6** (1990), 309-318.

[22] R. Villarreal, *Monomial algebras*, Marcel-Dekker (2001).

[23] M. S. Waterman, T. F. Smith, RNA secondary structure: a complete mathematical analysis, *Math. Biosci.* **42** (1978), 257–266.

[24] E. Westhof, L. Jaeger, L., RNA pseudoknots, *Curr. Opinion Struct. Biol.* **2** (1992) 327–333.

[25] M. Zuker, On finding all suboptimal foldings of an RNA molecule, *Science* **244** (1989), 48–52.

[26] M. Zuker, The use of dynamic programming algorithms in RNA secondary structure prediction, in *Mathematical methods for DNA sequences* (M. Waterman, ed.), CRC Press (1989), 159–184.

[27] M. Zuker, D. Sankoff, RNA secondary structures and their prediction, *Bull. Math. Biol.* **46** (1984), 591–621.