Journal of Mathematical Biology manuscript No.
(will be inserted by the editor)

Jaume Casasnovas · Joe Miro-Julia · Francesc Rosselló*

On the algebraic representation of RNA secondary structures with $G{\cdot}U$ pairs

Received: / Revised version:

Abstract. Magarshak *et al.* represented an RNA molecule as a complex vector $\underline{x}$ and an RNA secondary structure Γ as a complex matrix S_Γ in such a way that the molecule represented by $\underline{x}$ was compatible with the secondary structure Γ if and only if $S_\Gamma \circ \underline{x} = \underline{x}$. They only considered Watson-Crick base pairs and their representation cannot be extended to allow for $G \cdot U$ pairs. In this paper we study a generalization of Magarshak's representation that allows for these pairs, and in particular we provide a family of algebraic structures where that generalization can be carried out. We also show that this representation can be used to compare secondary structures, through transfer matrices which transform the representation of one secondary structure into the representation of the other.

1. Introduction

An RNA (ribonucleic acid) molecule can be viewed as a chain of ribonucleotides. Each ribonucleotide is characterized by the base attached to it, which can be either adenine (A), cytosine (C), guanine (G) or uracil (U). An RNA molecule is uniquely determined by the sequence of bases along its chain, and it has a definite orientation. Thus, mathematically, an RNA molecule with n nucleotides (or of *length* n) can be viewed as a word $b_1 b_2 \dots b_n \in \{A, C, G, U\}^+$, where each b_i stands for the ith ribonucleotide, or rather the corresponding base, of the molecule.

In the cell and *in vitro*, an RNA molecule folds into a three-dimensional structure that determines its biochemical activity. This three-dimensional structure is held together by weak interactions called *hydrogen bonds* between pairs of non-consecutive[1] bases. Most of these bonds form between *Watson-Crick complementary bases*, i.e., between A and U and between C

* Corresponding author.

Departament de Matemàtiques i Informàtica, Universitat de les Illes Balears,
07071 Palma de Mallorca (Spain)
e-mail: {jaumec,joe,cesc}@ipc4.uib.es

Key words: RNA secondary structure, algebra, finite field

[1] Actually, a hydrogen bond can only form between bases that are at least three ribonucleotides apart in the chain, but we shall not take this restriction into account here.

and G, but a far from negligible amount of bonds also form between other pairs of bases, as for instance the $G{\cdot}U$ wobble pairs.

As different levels of precision are suitable for different problems, we can sometimes forget about the detailed description of the three-dimensional structure of an RNA molecule and simply focus our attention on its *secondary structure*: the set of its base pairs. A restriction must be added to the definition of secondary structure: if two bases b_i and b_j are paired, then neither b_i or b_j can bond with any other base. This restriction is called the *unique bonds condition*.

So, roughly, a secondary structure of an RNA molecule of length n can be represented as an undirected graph Γ with set of nodes $\{1,\ldots,n\}$ and whose arcs (called *contacts*) satisfy the following two conditions: there is no contact between two consecutive nodes, and every node is involved in at most one contact. We shall consistently call such a graph an *RNA secondary structure*.

During the last years, there has been an increasing interest in the representation, at different levels of *graining*, of RNA and other biomolecules' secondary structures [9,11]. In this line of research, Y. Magarshak and his coworkers [4,7] represented an RNA molecule of length n as an n-dimensional vertical vector of complex numbers $\underline{x}$ and an RNA secondary structure Γ with n nodes as an $n \times n$ complex matrix S_Γ, in such a way that the molecule represented by the vector $\underline{x}$ was compatible with the secondary structure Γ if and only if $S_\Gamma \circ \underline{x} = \underline{x}$. In that work, *compatible* meant that any contact in the secondary structure paired two Watson-Crick complementary bases in the molecule.

It is important to have an algebraic representation of secondary structures as matrices over fields, as this allows the creation of transfer matrices that capture the difference between two structures. Once you have these matrices, all the power of linear algebra is at your disposition to study such difference. These kind of comparisons are essential to ascertain evolutionary processes.

Reidys and Stadler [9] pointed out that Magarshak's formalism did not allow for $G{\cdot}U$ wobble pairs, thus corresponding to a non-realistic definition of compatibility between sequences and secondary structures. This is unfortunate, as these pairs have frequently high evolutionary conservation, together with unique chemical, structural, dynamic and ligand-binding properties [13]. Reidys and Stadler pointedly asked (*op. cit.*) whether there is a framework in which Magarshak's ideas can be extended such as to allow for a more general logic of base pairing.

It turns out that such a framework does exist, provided one is willing to use a general enough algebraic structure. In this paper we discuss this problem in detail in the case of RNA secondary structures where $G \cdot U$ wobble pairs are allowed; the solution can be easily generalized to allow for other non-Watson-Crick base pairings. The framework presented here allows the creation of transfer matrices, that show how a secondary structure can

transform into another, together with a convenient way of measuring the difference between secondary structures.

We are working on a general formalism, corresponding to contact structures not necessarily satisfying the unique bonds condition and having possibly different types of contacts. This general formalism can be used to model tertiary structures of RNA or proteins. It will be presented some time in the near future.

Acknowledgements. We thank the anonymous referees for their comments, which have led to a substantial improvement of this paper. This work has been partially supported by the Spanish DGES, grant BFM2000-1113-C02-01.

2. What do we mean by an algebraic representation of RNA secondary structures

Let us start by introducing the formal definition of RNA secondary structure to be used in this paper. From now on, we denote by $[n]$ the set $\{1, \ldots, n\}$ for every positive integer n.

Definition 1. *An* RNA secondary structure *is an undirected graph without multiple edges or self-loops $\Gamma = ([n], Q)$, for some $n \geq 1$, whose arcs $\{j, k\} \in Q$, called* contacts*, satisfy the following two conditions:*

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

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

Condition (i) translates the impossibility of a contact between two consecutive bases, while condition (ii) is the aforementioned *unique bonds condition*. 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* when it is not involved in any contact.

Notice that our abstract definition of secondary structure is not the usual one, as the latter forbids the existence of *pseudoknots*: pairs of contacts $\{i, j\}$ and $\{k, l\}$ such that $i < k < j < l$. Thus, our term "RNA secondary structure" should be taken as an abbreviation for "RNA secondary structure with arbitrary knots," and it should be clear that it corresponds rather to what in the literature on secondary structure modelling has been called *contact structures with unique bonds* [9] or *1-diagrams* [3].

In [4,7], and following a well-established tradition in spectral graph theory (see [1, p. 2]), Magarshak and his coworkers represented an RNA secondary structure $\Gamma = ([n], Q)$ as the $n \times n$ complex symmetric matrix

$$S_\Gamma = \begin{pmatrix} s_{1,1} & \cdots & s_{1,n} \\ \vdots & \ddots & \vdots \\ s_{n,1} & \cdots & s_{n,n} \end{pmatrix}$$

where

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

It is straightforward to prove that $S_\Gamma \circ S_\Gamma = \mathrm{Id}_n$, the $n \times n$ identity matrix.

On the other hand, they used the correspondence

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

to represent an RNA molecule $\underline{b} = b_1 b_2 \dots b_n$ as a complex vector

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

These representations had the following key property:

> for every RNA molecule $\underline{b} = b_1 b_2 \dots b_n$ of length n and for every RNA secondary structure $\Gamma = ([n], Q)$, if $\underline{x} \in \mathbb{C}^n$ is the complex vector representing $\underline{b}$, then $S_\Gamma \circ \underline{x} = \underline{x}$ if and only if $\underline{b}$ is compatible with Γ, in the sense that if $j{\cdot}k \in Q$, then either $\{b_j, b_k\} = \{A, U\}$ or $\{b_j, b_k\} = \{C, G\}$.

Our goal is to define a representation of RNA secondary structures and molecules in the same spirit, but so that the notion of compatibility of a molecule $\underline{b}$ with a structure Γ includes $G{\cdot}U$ wobble pairs, i.e., the possibility that if $j{\cdot}k \in Q$, then $\{b_j, b_k\}$ can also be $\{G, U\}$. To do so, our plan is to represent RNA secondary structures and molecules as matrices and vectors with entries in what we shall call a *qs-ring*.

Definition 2. *A* quasi-semiring *(a* qs-ring*, for short) is an algebraic structure* $(X, +, *, 0, 1)$ *in which* $(X, +, 0)$ *is a commutative monoid,* $(X, *, 1)$ *is a monoid, and* $0 * x = x * 0 = 0$ *for every* $x \in X$*. Such a qs-ring is* commutative *when* $*$ *is commutative.*

The product $A \circ B$ *of two matrices*

$$A = \begin{pmatrix} a_{1,1} & \dots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \dots & a_{n,m} \end{pmatrix} \quad B = \begin{pmatrix} b_{1,1} & \dots & b_{1,p} \\ \vdots & \ddots & \vdots \\ b_{m,1} & \dots & b_{m,p} \end{pmatrix}$$

with entries in a qs-ring $(X, +, *, 0, 1)$ *is the matrix*

$$A \circ B = \begin{pmatrix} c_{1,1} & \dots & c_{1,p} \\ \vdots & \ddots & \vdots \\ c_{n,1} & \dots & c_{n,p} \end{pmatrix}$$

with $c_{j,k} = a_{j,1} * b_{1,k} + a_{j,2} * b_{2,k} + \dots + a_{j,m} * b_{m,k}$ *for every* $j = 1, \dots, n$, $k = 1, \dots, p$.

Semirings (and in particular rings and fields) and bounded lattices are specific examples of qs-rings: Appendix A is a reminder of the definitions of these and other algebraic structures appearing in this paper.

The choice of qs-rings from the start as our target algebraic structure is a compromise between generality and our goals: we need two binary operations, $+$ and $*$, in order to use matrix products; it is useful that the sum $+$ be associative and commutative to compute the product of two matrices in a safe way; we want to have two elements 0 and 1 so that

$$(\overbrace{0,\dots,0}^{j-1},1,0,\dots,0)\circ\begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix}=x_j,$$

which motivates to take 0 and 1 as the neutral elements of $+$ and $*$, respectively, and to impose $0*x$ to be always 0; and, finally, we ask the product $*$ to be associative because this restriction turns out to do no harm.

Once chosen a suitable qs-ring and four suitable pairwise different elements a,c,g,u in it, we shall use the coding

$$A\mapsto a,\ C\mapsto c,\ G\mapsto g,\ U\mapsto u$$

and we shall represent an RNA molecule $\underline{b}=b_1\dots b_n$ of length n by an n-dimensional vertical vector $\underline{x}$ with entries in $\{a,c,g,u\}$. Also, we will represent an RNA secondary structure $\Gamma=([n],Q)$ as the $n\times n$ symmetric matrix

$$S_\Gamma=\begin{pmatrix}s_{1,1}&\dots&s_{1,n}\\ \vdots&\ddots&\vdots\\ s_{n,1}&\dots&s_{n,n}\end{pmatrix}$$

where

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

for some suitable elements α,β of the qs-ring.

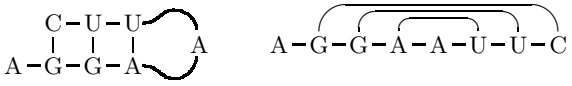

Fig. 1. Short RNA strand with its secondary structure shown graphically and schematically.

As an example, let us consider the seecondary structure of the short RNA strand shown in Figure 1. The structure of this strand of length 8 will

be represented by the following 8×8 matrix:

$$S_\Gamma = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 & 0 & 0 & 0 & \beta \\ 0 & 0 & \alpha & 0 & 0 & 0 & \beta & 0 \\ 0 & 0 & 0 & \alpha & 0 & \beta & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta & 0 & \alpha & 0 & 0 \\ 0 & 0 & \beta & 0 & 0 & 0 & \alpha & 0 \\ 0 & \beta & 0 & 0 & 0 & 0 & 0 & \alpha \end{pmatrix}$$

So, our goal becomes to look for a qs-ring $\mathbf{M} = (M, +, *, 0, 1)$ and six elements $\alpha, \beta, a, c, g, u$ in it such that, with the given coding and matrix representation, $S_\Gamma \circ \underline{x} = \underline{x}$ holds if and only if $\underline{b}$ is compatible with Γ, now in the sense that if $j \cdot k \in Q$, then $\{b_j, b_k\} = \{A, U\}, \{C, G\}$ or $\{G, U\}$.

It is straightforward to check that this equivalence holds if and only if the following conditions are satisfied in $\mathbf{M}$:

(1) $\alpha * a + \beta * u = a$
(2) $\alpha * u + \beta * a = u$
(3) $\alpha * u + \beta * g = u$
(4) $\alpha * g + \beta * u = g$
(5) $\alpha * g + \beta * c = g$
(6) $\alpha * c + \beta * g = c$
(7) $\alpha * a + \beta * g \neq a$ or $\alpha * g + \beta * a \neq g$
(8) $\alpha * a + \beta * c \neq a$ or $\alpha * c + \beta * a \neq c$
(9) $\alpha * u + \beta * c \neq u$ or $\alpha * c + \beta * u \neq c$
(10) $\alpha * a + \beta * a \neq a$
(11) $\alpha * u + \beta * u \neq u$
(12) $\alpha * g + \beta * g \neq g$
(13) $\alpha * c + \beta * c \neq c$

Pairs of conditions (1) and (2), (3) and (4), and (5) and (6) allow, respectively, the base pairings $A{\cdot}U$, $G{\cdot}U$, and $G{\cdot}C$, while the remaining conditions (7) through (13) forbid all other base pairings.

Recall that, in Magarshak's representation, α and β were taken to be 0 and -1, respectively. This choice is incompatible with our conditions, as the following easy lemma shows.

Lemma 3. *The elements α and β must be different from 0 as well as different from each other.*

Proof. If $\alpha = 0$, then equalities (1) and (4) imply that $a = \beta * u = g$. If $\beta = 0$ then, for instance,

$$\alpha * a + \beta * c = \alpha * a = \alpha * a + \beta * u = a$$
$$\alpha * c + \beta * a = \alpha * g = \alpha * g + \beta * u = c,$$

against condition (8). Finally, if $\alpha = \beta$, then $a = \alpha{*}a{+}\beta{*}u = \beta{*}a{+}\alpha{*}u = u$. □

3. What cannot be done

Let us discuss now the possibility of using algebraic structures richer (in properties) than bare qs-rings in our representation of RNA secondary structures. To begin with, the next two results exclude rings and bounded lattices.

Proposition 4. *Let* $\mathbf{M} = (M, +, *, 0, 1)$ *be a qs-ring and let* $\alpha, \beta, a, c, g, u$ *be any elements in it satisfying equalities (1) through (6) above. If* $(M, +, 0)$ *is a group, then they do not satisfy condition (8).*

Proof. Assume that $(M, +, 0)$ is a group. From equalities (2) and (3) we obtain that $\beta * a = \beta * g$, while equalities (4) and (5) imply that $\beta * u = \beta * c$. Then, using (1), (6) and these two last equalities, we deduce that

$$\alpha * a + \beta * c = \alpha * a + \beta * u = a$$
$$\alpha * c + \beta * a = \alpha * c + \beta * g = c$$

as we claimed. □

Proposition 5. *Let* $\mathbf{M} = (M, +, *, 0, 1)$ *be a qs-ring and let* $\alpha, \beta, a, c, g, u$ *be any elements in it satisfying equalities (1) through (6) above. If* $\mathbf{M}$ *is a bounded lattice, then they do not satisfy condition (8).*

Proof. Let $\leq$ be the partial order on M associated to this lattice. Then,

(1) implies $\beta * u \leq a$	(2) implies $\beta * a \leq u$
(3) implies $\beta * g \leq u$	(4) implies $\beta * u \leq g$
(5) implies $\beta * c \leq g$	(6) implies $\beta * g \leq c$

and thus

$$\beta * a \leq u,\ \beta * u \leq \alpha * g,\ \beta * g \leq u * c,\ \beta * c \leq g.$$

Then

$$\beta * a \leq \beta * u \leq \beta * \alpha * g \leq \beta * a$$

implies

$$\beta * a = \beta * u \leq \beta * g.$$

In a similar way,

$$\beta * c \leq \beta * g \leq \beta * u * c \leq \beta * c$$

implies

$$\beta * c = \beta * g \leq \beta * u.$$

Combining these equalities and inequalities we obtain

$$\beta * a = \beta * u = \beta * g = \beta * c,$$

from where, using (1) and (6), we finally deduce

$$\alpha * a + \beta * c = \alpha * a + \beta * u = a$$
$$\alpha * c + \beta * a = \alpha * c + \beta * g = c$$

as we claimed. □

After these first two negative results, let us further analyze our setting, to impose some more conditions on our qs-ring $\mathbf{M}$.

First, although $\mathbf{M} = (M, +, *, 0, 1)$ cannot be a ring, we ask it to extend a *field* $\mathbf{M}_0 = (M_0, +, *, 0, 1)$ and that the elements α, β belong to this field. This is done so that the operations only involving matrices S_Γ are safely carried out in this field (see Sect. 5), forgetting about the limitations of the qs-ring, which would only appear when RNA molecules are also considered.

Second, we can also impose, as in Magarshak's approach, that $S_\Gamma \circ S_\Gamma = \mathrm{Id}$. A simple computation shows that

$$S_\Gamma \circ S_\Gamma = \begin{pmatrix} a_{1,1} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \dots & a_{n,n} \end{pmatrix}$$

where

$$a_{j,j} = \begin{cases} 1 & \text{if } j \cdot l \notin Q \text{ for every } l \\ \alpha^2 + \beta^2 & \text{if } j \cdot l \in Q \text{ for some } l \end{cases} \quad a_{j,k} = \begin{cases} \alpha * \beta + \beta * \alpha & \text{if } j \cdot k \in Q \\ 0 & \text{otherwise} \end{cases}$$

where, of course, x^2 stands for $x * x$. Therefore, we require the following two extra conditions

$$(14) \qquad \alpha^2 + \beta^2 = 1$$
$$(15) \qquad \alpha * \beta + \beta * \alpha = 0$$

to hold in the field $\mathbf{M}_0$.

Lemma 6. *If a field $\mathbf{M}_0$ contains two elements $\alpha, \beta \neq 0$ such that $\alpha^2 + \beta^2 = 1$ and $\alpha * \beta + \beta * \alpha = 0$, then $\mathbf{M}_0$ has characteristic 2 and $\alpha + \beta = 1$.*

Proof. ¿From $\alpha^2 + \beta^2 = 1$ and $\alpha * \beta + \beta * \alpha = 0$ we deduce that $(\alpha + \beta)^2 = 1$. Then

$$(1 + (\alpha + \beta)) * (1 - (\alpha + \beta)) = 1 - (\alpha + \beta)^2 = 0$$

implies that $\alpha + \beta = \pm 1$. But then $\beta = \pm 1 - \alpha$ and

$$0 = \alpha * \beta + \beta * \alpha = 2\alpha(\pm 1 - \alpha).$$

If the field $\mathbf{M}_0$ is not of characteristic 2, then $\alpha = 0$ or $\alpha = \pm 1$, which contradicts the assumption $\alpha, \beta \neq 0$. Thus, $\mathbf{M}_0$ is a field of characteristic 2 and $\alpha + \beta = \pm 1 = 1$. □

Notice, in connection with this result, that if $\mathbf{M}_0$ is a field of characteristic 2 and $\alpha, \beta \in \mathbf{M}_0$ are such that $\alpha + \beta = 1$, then $\alpha^2 + \beta^2 = 1$, too.

Finally, we have still another negative result, which excludes semirings.

Proposition 7. *Let $\mathbf{M} = (M, +, *, 0, 1)$ be a qs-ring extending a field $\mathbf{M}_0$ and let $\alpha, \beta \in M_0$ and $a, c, g, u \in M$ be elements satisfying conditions (1) through (15) above. If the distributive laws*

$$x * (y + z) = x * y + x * z \text{ and } (y + z) * x = y * x + z * x$$

hold in $\mathbf{M}$ for every $x, y, z \in M$, then $a = u$.

Proof. Assume that $\mathbf{M}_0$ is a field (of characteristic 2) and that the operations in $\mathbf{M}$ satisfy the distributive laws. Multiplying both sides of $\alpha * a + \beta * u = a$ by the inverse β^{-1} of β in $\mathbf{M}_0$, we obtain that $(\beta^{-1} * \alpha) * a + u = \beta^{-1} * a$. Then

$$\begin{aligned} u &= (\beta^{-1} * \alpha + \beta^{-1} * \alpha) * a + u = \beta^{-1} * \alpha * a + \beta^{-1} * \alpha * a + u \\ &= \beta^{-1} * \alpha * a + \beta^{-1} * a = \beta^{-1} * (\alpha + 1) * a = \beta^{-1} * \beta * a = a \end{aligned}$$

as we claimed. □

4. A family of solutions

Summarizing the previous two sections, we are looking for qs-rings $\mathbf{M}$ extending fields $\mathbf{M}_0$ and containing six elements $\alpha, \beta, a, c, g, u \in M$ (with $\alpha, \beta \in M_0 - \{0\}$ different from each other and a, c, g, u pairwise different) satisfying conditions (1) through (15). We already know that $\mathbf{M}$ can be neither a semiring nor a bounded lattice, and that $\mathbf{M}_0$ must be of characteristic 2. In this section we show that every finite field $\mathbb{F}_{2^m}$ of characteristic 2 (except $\mathbb{F}_2$, which does not contain two elements different from 0) can be extended to a commutative qs-ring $\mathbf{M}_m$ solving our problem.

Let us quickly recall that $\mathbb{F}_2$ is simply the quotient ring $\mathbb{Z}/2\mathbb{Z}$ (with elements $0, 1$) and, for $m \geq 2$, the field $\mathbb{F}_{2^m}$ is commutative, has 2^m elements, and it is obtained as a quotient $\mathbb{F}_2[x]/\langle q_m(x)\rangle$, where $q_m(x)$ is a *primitive* irreducible polynomial of degree m over $\mathbb{F}_2$ (see [5, Chap. 3]). Recall moreover that $(\mathbb{F}_{2^m} - \{0\}, *, 1)$ is a *cyclic* group generated by the class of x modulo $q_m(x)$.

The extension of a field $\mathbb{F}_{2^m}$, $m \geq 2$, to a commutative qs-ring $\mathbf{M}_m$ containing four elements a, c, g, u that satisfy conditions (1) through (13) with respect to two suitable elements $\alpha, \beta \in \mathbb{F}_{2^m} - \{0\}$ such that $\alpha^2 + \beta^2 = 1$ (or, equivalently, such that $\alpha + \beta = 1$) is defined in the following way:

(1) We take α to be the class of the variable x in $\mathbb{F}_2[x]$ modulo the specific primitive polynomial $q_m(x)$ used to define $\mathbb{F}_{2^m}$.
(2) We take as β the element $\alpha + 1$ in $\mathbb{F}_{2^m}$. Since $(\mathbb{F}_{2^m} - \{0\}, *, 1)$ is a cyclic group with $2^m - 1$ elements generated by α, and since the polynomial $q_m(x)$ is irreducible, we have that $\beta = \alpha^{\ell_m}$ for some exponent $n \leq \ell_n \leq 2^m - 2$. The following table lists some pairs $(q_m(x), \ell_m)$.

m	$q_m(x)$	ℓ_m
2	$x^2 + x + 1$	2
3	$x^3 + x^2 + 1$	5
4	$x^4 + x + 1$	4
5	$x^5 + x^2 + 1$	18
6	$x^6 + x + 1$	6
7	$x^7 + x + 1$	7

(3) The carrier set M_m of the qs-ring $\mathbf{M}_m$ is the (disjoint) union of $\mathbb{F}_{2^m}$ and a finite set

$$X_m = \{a, c, g, u\} \cup \{a_j, c_j, g_j, u_j \mid j = 1, \ldots, 2^m - 2\} \cup \{z\};$$

the extra element z is actually only needed for $m = 2$, but for homogeneity we add it in all cases.

(4) The product $*$ on M_m is defined as follows:
- $*$ is taken to be commutative;
- $x * y = 0$ for every $x, y \in X_m$;
- the product of elements of $\mathbb{F}_{2^m}$ is the usual one;
- $0 * x = 0$ and $1 * x = x$ for every $x \in X_m$;
- the product of elements of X_m by α

$$\begin{aligned} X_m &\to X_m \\ x &\mapsto \alpha * x \end{aligned}$$

is given by the permutation on X_m of order $2^m - 1$

$$\begin{aligned}(a, a_1, a_2, \ldots, a_{2^m-2}) \cdot (c, c_1, c_2, \ldots, c_{2^m-2}) \\ \cdot (g, g_1, g_2, \ldots, g_{2^m-2}) \cdot (u, u_1, u_2, \ldots, u_{2^m-2}),\end{aligned}$$

i.e.,

$$\alpha * a = a_1,\ \alpha * a_1 = a_2, \ldots, \alpha * a_{2^m-3} = a_{2^m-2},\ \alpha * a_{2^m-2} = a$$

and similarly for c, g and u, and $\alpha * z = z$.
- for every $\lambda \in \mathbb{F}_{2^m} - \{0, 1, a\}$, if l is the unique exponent between 2 and $2^m - 2$ such that $\lambda = \alpha^l$, then

$$\lambda * x = \overbrace{\alpha * (\alpha * \cdots (\alpha}^{l} * x) \cdots) \text{ for every } x \in X_m;$$

in particular

$$\beta * a = a_{\ell_m},\ \beta * c = c_{\ell_m},\ \beta * g = g_{\ell_m},\ \beta * u = u_{\ell_m}.$$

(5) The sum $+$ on M_m is defined as follows:
- $+$ is taken to be commutative;
- the sum of elements of $\mathbb{F}_{2^m}$ is the usual one;
- $\lambda + x = x$ for every $\lambda \in \mathbb{F}_{2^m}$ and $x \in X_m$;
- conditions (1) through (6), together with the commutativity of $+$, yield

$$\begin{aligned} a_1 + u_{\ell_m} &= u_{\ell_m} + a_1 = a \\ u_1 + a_{\ell_m} &= a_{\ell_m} + u_1 = u \\ u_1 + g_{\ell_m} &= g_{\ell_m} + u_1 = u \\ g_1 + u_{\ell_m} &= u_{\ell_m} + g_1 = g \\ g_1 + c_{\ell_m} &= c_{\ell_m} + g_1 = g \\ c_1 + g_{\ell_m} &= g_{\ell_m} + c_1 = c \end{aligned}$$

- all other sums $x + y$ with $x, y \in X_m$ are defined to be equal to a fixed element in X_m different from

$$a, c, g, u, a_1, c_1, g_1, u_1, a_{\ell_m}, c_{\ell_m}, g_{\ell_m}, u_{\ell_m};$$

for homogeneity, we choose z as this element.

Theorem 8. *The algebraic structure* $\mathbf{M}_m = (M_m, +, *, 0, 1)$ *defined in this way is, for every* $m \geq 2$*, a qs-ring extending* $\mathbb{F}_{2^m}$*, and the elements* a, c, g, u*,* α, β *in it satisfy conditions (1) through (15).*

Proof. The only property that is not straightforward to prove is the associativity of the sum, but it is a direct consequence of the fact that "one-point completions preserve strong equations": see [12, Thm. 5.1]. □

Remark 9. These algebras can be easily modified in order to cope with other non-canonical base pairings. For instance, assume that we want to allow also sheared $G{\cdot}A$ base pairings, which are not uncommon [14]. Then, we must replace condition (7) by its negation

$$\alpha * a + \beta * g = a \quad \text{and} \quad \alpha * g + \beta * a = g.$$

The effect on $\mathbf{M}_m$ is that, instead of defining the sums $a_1 + g_{\ell_m}$, $g_{\ell_m} + a_1$, $g_1 + a_{\ell_m}$, and $a_{\ell_m} + g_1$ to be equal to z, we must redefine them as

$$a_1 + g_{\ell_m} = g_{\ell_m} + a_1 = a, \quad g_1 + a_{\ell_m} = a_{\ell_m} + g_1 = g$$

without changing anything else. The new algebraic structure on M_m defined in this way is still a qs-ring, and now, with the given coding and matrix representation, $S_\Gamma \circ \underline{x} = \underline{x}$ holds in this qs-ring if and only if every contact in the structure corresponds to a base pairing $A{\cdot}U$, $C{\cdot}G$, $G{\cdot}U$, or $G{\cdot}A$ in the molecule.

These algebras can also be easily modified in order to cope with extended sets of ribonucleic bases and base pairings. For instance, the base pairing among the synthetic bases xanthine and 2,6-diaminopyrimidine has been considered both in the biochemical [8] and in the secondary structure modelling [2] literature. In order to model RNA molecules as words over the alphabet $\{A, C, G, U, X, K\}$ (where now X stands for xanthine and K for 2,6-diaminopyrimidine) and to allow $A{\cdot}U$, $C{\cdot}G$, $G{\cdot}U$, and $X{\cdot}K$ base pairings, it is enough:

– to add new elements

$$\{x, k\} \cup \{x_j, k_j \mid j = 1, \ldots, 2^m - 2\}$$

to M_m, where x and k will, of course, encode the new bases X and K, respectively;

– to define the product of these new elements by α by means of the permutations

$$(x, x_1, x_2, \ldots, x_{2^m-2}), (k, k_1, k_2, \ldots, k_{2^m-2})$$

as we already did with the "old" elements in X_m;

– to define

$$x_1 + k_{\ell_m} = k_{\ell_m} + x_1 = x, \quad k_1 + x_{\ell_m} = x_{\ell_m} + k_1 = k$$

and all other new sums to be equal to z: this will impose that

$$\alpha * x + \beta * k = x, \quad \alpha * k + \beta * x = k,$$

and that no other relation of this kind is added.

The new algebraic structure defined in this way is again a qs-ring, and now, with the given coding and matrix representation, $S_\Gamma \circ \underline{x} = \underline{x}$ holds in this qs-ring if and only if every contact in the structure corresponds to an allowed base pairing in the molecule.

5. Transfer matrices

Let $\mathbb{F}_{2^m}$ be a finite field of characteristic 2 with $m \geq 2$, and let $\alpha, \beta \in \mathbb{F}_{2^m}$ be the elements used in the last section to define the matrices S_Γ. To simplify the notations, throughout this section we shall omit the symbol $*$ to denote the product in this field, thus writing xy instead of $x * y$; we shall assume that all RNA secondary structures appearing in this section have the same number n of nodes; and given an RNA secondary structure $\Gamma = ([n], Q)$ and a node $i \in [n]$, we shall denote $Q(i) = \{j \in [n] \mid i \cdot j \in Q\}$.

Definition 10. *The* transfer matrix T_{Γ_1,Γ_2} *of two RNA secondary structures* Γ_1 *and* Γ_2 *is* $T_{\Gamma_1,\Gamma_2} = S_{\Gamma_2} \circ S_{\Gamma_1}$.

Since $S_{\Gamma_2} = (S_{\Gamma_2} \circ S_{\Gamma_1}) \circ S_{\Gamma_1}$, this *transfer matrix* T_{Γ_1,Γ_2} converts the matrix representation of Γ_1 into the representation of Γ_2:

$$T_{\Gamma_1,\Gamma_2} \circ S_{\Gamma_1} = S_{\Gamma_2}.$$

The next lemma collects some basic compositional properties of these transfer matrices.

Lemma 11. *Let* $\Gamma_1, \Gamma_2, \Gamma_3$ *be any RNA secondary structures.*
i) $T_{\Gamma_1,\Gamma_1} = \mathrm{Id}$.
ii) $T_{\Gamma_1,\Gamma_3} = T_{\Gamma_2,\Gamma_3} \circ T_{\Gamma_1,\Gamma_2}$.
iii) $T_{\Gamma_2,\Gamma_1} = (T_{\Gamma_1,\Gamma_2})^{-1} = (T_{\Gamma_1,\Gamma_2})^t$. *In particular,* $\det(T_{\Gamma_1,\Gamma_2}) = 1$. □

An elementary but tedious computation (which we leave to the reader) produces the following explicit description of the transfer matrix T_{Γ_1,Γ_2}.

Proposition 12. *Let* $\Gamma_1 = ([n], Q_1)$ *and* $\Gamma_2 = ([n], Q_2)$ *be two RNA secondary structures. Then*

$$T_{\Gamma_1,\Gamma_2} = \begin{pmatrix} t_{1,1} & \dots & t_{1,n} \\ \vdots & \ddots & \vdots \\ t_{n,1} & \dots & t_{n,n} \end{pmatrix}$$

where, for every $i = 1, \ldots, n$,

$$t_{i,i} = \begin{cases} 1 & \text{if } Q_1(i) = Q_2(i) \\ \alpha & \text{if } Q_1(i) = \emptyset \text{ and } Q_2(i) \neq \emptyset \\ & \quad \text{or } Q_1(i) \neq \emptyset \text{ and } Q_2(i) = \emptyset \\ \alpha^2 & \text{if } Q_1(i), Q_2(i) \neq \emptyset \text{ and } Q_1(i) \neq Q_2(i) \end{cases}$$

and, for every $1 \leq i \neq j \leq n$

$$t_{i,j} = \begin{cases} \beta^2 & \text{if } Q_1(j) = Q_2(i) \neq \emptyset \\ \beta & \text{if } Q_2(i) = \{j\} \text{ and } Q_1(j) = \emptyset \\ & \quad \text{or } Q_1(j) = \{i\} \text{ and } Q_2(i) = \emptyset \\ \alpha\beta & \text{if } Q_2(i) = \{j\} \text{ and } Q_1(j) \neq \emptyset, \{i\} \\ & \quad \text{or } Q_1(j) = \{i\} \text{ and } Q_2(i) \neq \emptyset, \{j\} \\ 0 & \text{in any other case} \end{cases}$$

□

Notice in particular that T_{Γ_1,Γ_2} does not depend on Γ_1 and Γ_2, but rather on the differences between them.

Example 13. Let $\Gamma_1 = ([19], Q_1)$ and $\Gamma_2 = ([19], Q_2)$ be the RNA secondary structures defined by the sets of contacts

$$Q_1 = \{1{\cdot}3, 5{\cdot}7, 9{\cdot}11, 13{\cdot}15, 17{\cdot}19\}$$
$$Q_2 = \{3{\cdot}5, 8{\cdot}16, 9{\cdot}15, 11{\cdot}13, 17{\cdot}19\};$$

see Fig. 2. Then:

$$S_{\Gamma_1} = \begin{pmatrix}
\alpha & 0 & \beta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\beta & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \alpha & 0 & \beta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \beta & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \beta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \beta & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta & 0 & \alpha & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \beta \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta & 0 & \alpha
\end{pmatrix}$$

$$S_{\Gamma_2} = \begin{pmatrix}
1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&\alpha&0&\beta&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&\beta&0&\alpha&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&\alpha&0&0&0&0&0&0&\beta&0&0&0&0\\
0&0&0&0&0&0&0&0&\alpha&0&0&0&0&\beta&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&\alpha&0&\beta&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&\beta&0&\alpha&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&\beta&0&0&0&0&0&\alpha&0&0&0&0&0\\
0&0&0&0&0&0&0&\beta&0&0&0&0&0&0&0&\alpha&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&\alpha&0&\beta&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&\beta&0&\alpha&0
\end{pmatrix}$$

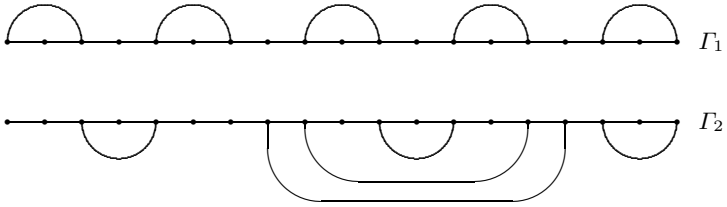

Fig. 2. The secondary structures in Example 13.

Computing their product $S_{\Gamma_2} \circ S_{\Gamma_1}$ we obtain the transfer matrix

$$T_{\Gamma_1,\Gamma_2} = \begin{pmatrix}
\alpha&0&\beta&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
\alpha\beta&0&\alpha^2&0&\alpha\beta&0&\beta^2&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
\beta^2&0&\alpha\beta&0&\alpha^2&0&\alpha\beta&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&\beta&0&\alpha&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&\alpha&0&0&0&0&0&0&0&\beta&0&0&0&0\\
0&0&0&0&0&0&0&0&\alpha^2&0&\alpha\beta&0&\beta^2&0&\alpha\beta&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&\alpha\beta&0&\alpha^2&0&\alpha\beta&0&\beta^2&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&\beta^2&0&\alpha\beta&0&\alpha^2&0&\alpha\beta&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&\alpha\beta&0&\beta^2&0&\alpha\beta&0&\alpha^2&0&0&0&0&0\\
0&0&0&0&0&0&0&\beta&0&0&0&0&0&0&0&\alpha&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1
\end{pmatrix}$$

The transfer matrices T_{Γ_1,Γ_2} can be used to compare RNA secondary structures. For instance, in [9] (and for Magarshak's matrices, considering them as real-valued) Reidys and Stadler proposed to measure the distance between two secondary structures Γ_1 and Γ_2 by means of the value $\|T_{\Gamma_1,\Gamma_2}\|$ for an arbitrary matrix norm $\|\cdot\|$. Using matrix norms does not make sense in our setting, because our matrices have their entries in a field of characteristic 2. Anyway, there are other values computed from our transfer matrices that can be used to measure the difference between two secondary structures. More specifically, we shall use the following two values.

Definition 14. *Given two RNA secondary structures Γ_1 and Γ_2, let*

- *$\delta(\Gamma_1,\Gamma_2)$ be the product of the elements of the main diagonal of T_{Γ_1,Γ_2};*
- *$\rho(\Gamma_1,\Gamma_2)$ be the rank of $T_{\Gamma_1,\Gamma_2}+\mathrm{Id}$.*

Since $T_{\Gamma_2,\Gamma_1}=(T_{\Gamma_1,\Gamma_2})^t$, and thus $T_{\Gamma_2,\Gamma_1}+\mathrm{Id}=(T_{\Gamma_1,\Gamma_2}+\mathrm{Id})^t$, it is clear that $\delta(\Gamma_1,\Gamma_2)=\delta(\Gamma_2,\Gamma_1)$ and $\rho(\Gamma_1,\Gamma_2)=\rho(\Gamma_2,\Gamma_1)$.

Example 15. If Γ_1 and Γ_2 are the secondary structures given in Example 13, then $\delta(\Gamma_1,\Gamma_2)=\alpha^{16}$. As far as $\rho(\Gamma_1,\Gamma_2)$ goes, it is the rank of the following matrix (where we use that $1+1=0$, $\alpha+1=\beta$, and $\alpha^2+1=\beta^2$):

$$T_{\Gamma_1,\Gamma_2}+\mathrm{Id}=\begin{pmatrix}
\beta & 0 & \beta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\alpha\beta & 0 & \beta^2 & 0 & \alpha\beta & 0 & \beta^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\beta^2 & 0 & \alpha\beta & 0 & \beta^2 & 0 & \alpha\beta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & \beta & 0 & \beta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta^2 & 0 & \alpha\beta & 0 & \beta^2 & 0 & \alpha\beta & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha\beta & 0 & \beta^2 & 0 & \alpha\beta & 0 & \beta^2 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta^2 & 0 & \alpha\beta & 0 & \beta^2 & 0 & \alpha\beta & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha\beta & 0 & \beta^2 & 0 & \alpha\beta & 0 & \beta^2 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

Removing rows and columns containing only zeroes, and extracting a common factor β from each one of the remaining rows, we obtain the following

matrix with the same rank:

$$\begin{pmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\alpha & \beta & \alpha & \beta & 0 & 0 & 0 & 0 & 0 & 0 \\
\beta & \alpha & \beta & \alpha & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & \beta & \alpha & \beta & \alpha & 0 \\
0 & 0 & 0 & 0 & 0 & \alpha & \beta & \alpha & \beta & 0 \\
0 & 0 & 0 & 0 & 0 & \beta & \alpha & \beta & \alpha & 0 \\
0 & 0 & 0 & 0 & 0 & \alpha & \beta & \alpha & \beta & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}$$

Now we find several pairs of rows where we can apply the following transformation, or a similar one (in the first and second steps, we use again that $\alpha + \beta = 1$):

$$\begin{pmatrix} \alpha & \beta & \alpha & \beta \\ \beta & \alpha & \beta & \alpha \end{pmatrix} \Rightarrow \begin{pmatrix} \alpha & \beta & \alpha & \beta \\ 1 & 1 & 1 & 1 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

Using them, we finally obtain the following matrix with the same rank:

$$\begin{pmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}$$

and now a simple computation shows that this rank is 6.

Our last two theorems show how the closeness of two RNA secondary structures of the same length can be measured by means of these two values, and in particular they explain the figures $\delta(\Gamma_1, \Gamma_2) = \alpha^{16}$ and $\rho(\Gamma_1, \Gamma_2) = 6$ obtained in the last example. Their proofs are simple, but long and technically involved, and we delay them until Appendix B at the end of this paper.

Let $|Q_1 \triangle Q_2|$ denote in the sequel the cardinal of the symmetric difference of the sets of contacts of the secondary structures Γ_1 and Γ_2.

Theorem 16. *For every two RNA secondary structures $\Gamma_1 = ([n], Q_1)$ and $\Gamma_2 = ([n], Q_2)$,*

$$\delta(\Gamma_1, \Gamma_2) = \alpha^{2|Q_1 \triangle Q_2|}.$$

Notice that if we choose the exponent m in $\mathbb{F}_{2^m}$ large enough (actually, such that $n < 2^m - 1$), then $\delta(\Gamma_1, \Gamma_2) = 1$ if and only if $\Gamma_1 = \Gamma_2$. Thus, in this case, $|Q_1 \triangle Q_2|$ can be obtained as $\frac{1}{2}\log_\alpha \delta(\Gamma_1, \Gamma_2)$ in $\mathbb{F}_{2^m}$.

To give an explicit description of the value of $\rho(\Gamma_1, \Gamma_2)$, we need to introduce a further notion, extracted from [10, § 4]. Given two secondary structures $\Gamma_1 = ([n], Q_1)$ and $\Gamma_2 = ([n], Q_2)$, a *cyclic orbit* for Γ_1 and Γ_2 is a subset $\{i_1, i_2, \ldots, i_{2m}\} \subseteq [n]$, $m \geq 1$, such that either

$$i_1{\cdot}i_2, i_3{\cdot}i_4, \ldots, i_{2m-1}{\cdot}i_{2m} \in Q_1 \text{ and } i_2{\cdot}i_3, \ldots, i_{2m-2}{\cdot}i_{2m-1}, i_{2m}{\cdot}i_1 \in Q_2$$

or the other way round

$$i_1{\cdot}i_2, i_3{\cdot}i_4, \ldots, i_{2m-1}{\cdot}i_{2m} \in Q_2 \text{ and } i_2{\cdot}i_3, \ldots, i_{2m-2}{\cdot}i_{2m-1}, i_{2m}{\cdot}i_1 \in Q_1;$$

see Fig. 3. In particular, $\{i_1, i_2\}$ is a cyclic orbit when $i_1{\cdot}i_2 \in Q_1 \cap Q_2$. The unique bonds condition implies that a cyclic orbit must have an even number of elements (it can also be deduced by group-theoretical considerations: see the proof of [10, Thm. 5]). Let Ω be the number of cyclic orbits for Γ_1 and Γ_2 with cardinal at least 4. Now, we have the following result.

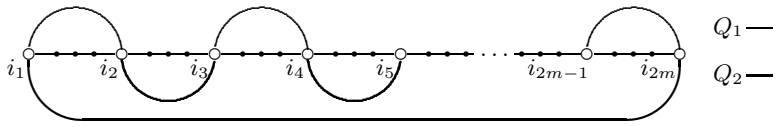

Fig. 3. A cyclic orbit.

Theorem 17. *For every two RNA secondary structures $\Gamma_1 = ([n], Q_1)$ and $\Gamma_2 = ([n], Q_2)$,*

$$\rho(\Gamma_1, \Gamma_2) = |Q_1 \triangle Q_2| - 2\Omega.$$

Notice that $\rho(\Gamma_1, \Gamma_2)$ is also equal to $|Q_1| + |Q_2|$ minus twice the number of *all* cyclic orbits for Γ_1 and Γ_2. This can be used to prove that $\rho(\Gamma_1, \Gamma_2)$ is exactly the value of the metric defined in [9, Thm. 4] applied to the pair (Γ_1, Γ_2).

References

1. F. R. K. Chung, *Spectral Graph Theory*. CBMS vol. 92, AMS (1994).
2. W. Fontana, D. A. M. Konings, P. F. Stadler, P. Schuster, Statistics of RNA Secondary Structures. *Biopolymers* **33** (1993), 1389–1404.
3. C. Haslinger, P. F. Stadler, RNA Structures with Pseudo-Knots: Graph-theoretical, Combinatorial, and Statistical Properties. *Bulletin of Mathematical Biology* **61** (1999), 437–467.

4. A. Kister, Y. Magarshak, J. Malinsky, The theoretical analysis of the process of RNA molecule self-assembly. *BioSystems* **30** (1993), 31–48.
5. R. Lidl, G. Pilz, *Applied Abstract Algebra*. Undergraduate Texts in Mathematics, Springer-Verlag (1984).
6. Y. Magarshak, Quaternion representation of RNA sequences and tertiary structures. *BioSystems* **30**, (1993) 21–29.
7. Y. Magarshak, C. J. Benham, An algebraic representation of RNA secondary structures. *J. of Biomolecular Structures & Dynamics* **10** (1992), 465–488.
8. J. A. Piccirilli, T. Krauch, S. E. Moroney, S. A. Benner, Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* **343** (1990), 33-37.
9. C. Reidys, P. F. Stadler, Bio-molecular shapes and algebraic structures. *Computers & Chemistry* **20** (1996), 85–94.
10. C. Reidys, P. F. Stadler, P. Schuster, Generic Properties of Combinatory Maps: Neural Networks of RNA Secondary Structures. *Bulletin of Mathematical Biology* **59** (1997), 339–397.
11. 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, editors), Marcel Dekker (in press). See also Univ. Wien TBI Preprint No. pks-99-012 (1999).
12. B. Staruch, B. Staruch, Strong regular varieties of partial algebras. *Algebra Universalis* **31** (1994), 157–176.
13. G. Varani, W. H. McClain, The *GU* wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Reports* **1** (2000), 18–23.
14. E. Westhof, V. Fritsch, RNA folding: beyond Watson-Crick pairs. *Structure with Folding & Design* **8** (2000), R55–R65.

Appendix A: A glossary of algebraic structures

For the convenience of the reader, in this appendix we remind the definitions of the algebraic structures mentioned in the main body of the paper.

Let X be a non-empty set. A binary operation $*$ on X is

- *commutative*, when $x * y = y * x$ for every $x, y \in X$;
- *associative*, when $(x * y) * z = x * (y * z)$ for every $x, y, z \in X$.

An element $e \in X$ is a *neutral element* for a binary operation $*$ when $x * e = x = e * x$ for every $x \in X$. Two binary operations $+$ and $*$ on X satisfy the *distributive law* when

$$x * (y + z) = x * z + y * z, \quad (y + z) * x = y * x + z * x, \qquad \text{for every } x, y, z \in X.$$

A *semigroup* is an algebraic structure $(X, *)$ consisting of a non-empty set X and an associative binary operation $*$ on it. A *monoid* is an algebraic structure $(X, *, e)$ such that $(X, *)$ is a semigroup and e is a neutral element for $*$. A *group* is a monoid $(X, *, e)$ such that, for every $x \in X$, there exists an element $x' \in X$ such that $x * x' = x' * x = e$. A semigroup, monoid or group is said to be *commutative* when its binary operation $*$ is commutative.

A *semiring* is an algebraic structure $(X, +, *, 0, 1)$ such that $(X, +, 0)$ is a commutative monoid, $(X, *, 1)$ is a monoid, and $*$ distributes over $+$. A *ring* is a semiring $(X, +, *, 0, 1)$ such that $(X, +, 0)$ is, moreover, a commutative group. A *field* is a ring $(X, +, *, 0, 1)$ such that, for every $x \in X - \{0\}$, there exists an element $x' \in X$ such that $x * x' = x' * x = 1$. A semiring, ring or field $(X, +, *, 0, 1)$ is said to be *commutative* when the binary operation $*$ is commutative.

A field $(X, +, *, 0, 1)$ is said to have *characteristic* n, for some natural number $n \geq 1$, when $\overbrace{1 + \cdots + 1}^{n} = 0$ and $\overbrace{1 + \cdots + 1}^{m} \neq 0$ for every $1 \leq m < n$: in this case, n must be a prime number. If a field does not have characteristic n for any $n \geq 1$, then it is said to have *characteristic 0*. In a field of characteristic $n \geq 1$ we have that $n * x = \overbrace{1 + \cdots + 1}^{n} = 0$ for every element x, while if the field does not have characteristic n, for some prime number n, then $n * x = 0$ implies $x = 0$.

A *lattice* is an algebraic structure $(X, +, *)$ such that both $(X, +)$ and $(X, *)$ are commutative semigroups, and moreover they satisfy the following two properties:

- *idempotency*: $x + x = x$ and $x * x = x$ for every $x \in X$;
- *absortion*: $x + (x * x) = x$ and $x * (x + x)$ for every $x \in X$.

Equivalently, $(X, +, *)$ is a lattice whenever the relation $\leq$ on X defined by

$$x \leq y \Longleftrightarrow x * y = x$$

(or, equivalently, $x \leq y \Leftrightarrow x + y = y$) is a partial order such that every pair of elements x, y has a supremum and an infimum, given, respectively, by $x + y$ and $x * y$.

A *bounded lattice* is an algebraic structure $(X, +, *, 0, 1)$ such that $(X, +, *)$ is a lattice and $x * 0 = 0$ and $x + 1 = 1$: this is equivalent to say that the elements 0 and 1 are, respectively, the smallest and the greatest elements of X with respect to the partial order $\leq$ defined above.

Appendix B: Proof of Theorems 16 and 17

We begin by framing the definition of cyclic orbit given in Sect. 5 by giving the general definition of an orbit for a pair of secondary structures. As its name hints, the origin of this concept lies in a certain group-theoretical interpretation of secondary structures that can be found, for instance, in [2,9].

Given two secondary structures $\Gamma_1 = ([n], Q_1)$ and $\Gamma_2 = ([n], Q_2)$, an *orbit* for Γ_1 and Γ_2 is a subset $\{i_1, i_2, \ldots, i_m\} \subseteq [n]$, $m \geq 1$, such that

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

and maximal with this property (i.e., such that any other contact in $Q_1 \cup Q_2$ involving i_1 or i_m 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 either

$$i_1 \cdot i_2, i_3 \cdot i_4, \ldots, \in Q_1 \text{ and } i_2 \cdot i_3, i_4 \cdot i_5, \ldots \in Q_2$$

or

$$i_1 \cdot i_2, i_3 \cdot i_4, \ldots, \in Q_2 \text{ and } i_2 \cdot i_3, i_4 \cdot i_5, \ldots \in Q_1.$$

It is clear from the maximality of the orbits and the unique bonds condition that two orbits for Γ_1 and Γ_2 are always disjoint.

An orbit $\{i_1, i_2, \ldots, i_m\}$ is said to be *linear* when:

- $m = 1$, and then this means that i_1 is isolated in both Γ_1 and Γ_2; or
- $m = 2$ and $i_1 \cdot i_2 \in Q_1 \triangle Q_2$; or
- $m \geq 3$ and $i_1 \cdot i_m \notin Q_1 \cup Q_2$.

An orbit $\{i_1, i_2, \ldots, i_m\}$ is said to be *cyclic* when:

- $m = 2$ and $i_1 \cdot i_2 \in Q_1 \cap Q_2$; or
- $m \geq 3$ and $i_1 \cdot i_m \in Q_1 \cup Q_2$.

As we already mentioned in Sect. 5, the cardinal of a cyclic orbit is always even.

For instance, the orbits for the secondary structures given in Example 13 are

$$\{8,10,12,14\},\{17,19\},\{1,3,5,7\},\{7,15\},\{2\},\{4\},\{6\},\{9\},\{11\},\{13\},\{18\};$$

the first two ones are cyclic and the other ones are linear.

Let now Γ_1 and Γ_2 be two arbitrary RNA secondary structures with n nodes. The orbits for Γ_1 and Γ_2 allow the decomposition of T_{Γ_1,Γ_2} into disjoint minors. Indeed, since every $i \in [n]$ belongs to one, and only one, orbit for Γ_1 and Γ_2, and since the rank of a matrix and the product of the elements of its main diagonal do not change under permutations of rows and columns, provided the same permutation is applied simultaneously to both rows and columns, when computing $\delta(\Gamma_1,\Gamma_2)$ and $\rho(\Gamma_1,\Gamma_2)$ we may assume without any loss of generality that the elements of each orbit are consecutive in $[n]$. It is clear then that the product of the elements of the main diagonal of T_{Γ_1,Γ_2} can be computed as the product of the products of the elements of the main diagonals of the minors defined by the orbits, and that the rank of $T_{\Gamma_1,\Gamma_2} + \mathrm{Id}$ can be computed as the sum of the ranks of the minors defined by the orbits.

Any linear orbit consisting of a singleton $\{i\}$ appears in the transfer matrix T_{Γ_1,Γ_2} as the ith row and column with a 1 on the main diagonal and the other entries equal to 0. Thus, it contributes a factor 1 to $\delta(\Gamma_1,\Gamma_2)$ and, since when we add up the identity matrix to the transfer matrix the 1 on the diagonal is cancelled, it contributes nothing to $\rho(\Gamma_1,\Gamma_2)$.

Any linear orbit with $p \geq 2$ elements corresponds to p elements in $Q_1 \triangle Q_2$, and it turns out to contribute a factor α^{2p} to $\delta(\Gamma_1,\Gamma_2)$ and to add up p to the $\rho(\Gamma_1,\Gamma_2)$. Let us show it in one case: we leave the other cases to the reader (special cases of two of them have appeared in the computation carried over in Example 15).

Assume that $\{i, i+1, i+2, \ldots, i+2m\}$ is a linear orbit with $i\cdot(i+1), (i+2)\cdot(i+3), \ldots, (i+2m-2)\cdot(i+2m-1) \in Q_1$ and $(i+1)\cdot(i+2), \ldots, (i+2m-3)\cdot(i+2m-2), (i+2m-1)\cdot(i+2m) \in Q_2$. A simple computation shows that the minor of T_{Γ_1,Γ_2} corresponding to rows and columns between the ith one and the $(i+2m)$th one is the following:

$$\begin{pmatrix}
\alpha & \beta & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\
\alpha\beta & \alpha^2 & \alpha\beta & \beta^2 & 0 & 0 & \ldots & 0 & 0 & 0 \\
\beta^2 & \alpha\beta & \alpha^2 & \alpha\beta & 0 & 0 & \ldots & 0 & 0 & 0 \\
0 & 0 & \alpha\beta & \alpha^2 & \alpha\beta & \beta^2 & \ldots & 0 & 0 & 0 \\
0 & 0 & \beta^2 & \alpha\beta & \alpha^2 & \alpha\beta & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & \ldots & \alpha\beta & \alpha^2 & \beta \\
0 & 0 & 0 & 0 & 0 & 0 & \ldots & \beta^2 & \alpha\beta & \alpha
\end{pmatrix}$$

and therefore the product of the entries in its main diagonal is $\alpha^{2(2m+1)}$,i.e., α raised to twice the cardinal of the orbit, as we claimed.

Now, to compute $\rho(\Gamma_1,\Gamma_2)$, we add 1 to the elements of the main diagonal of this minor and (using that $\alpha + 1 = \beta$ and $\alpha^2 + 1 = \beta^2$) we obtain the following matrix:

$$\begin{pmatrix}
\beta & \beta & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\
\alpha\beta & \beta^2 & \alpha\beta & \beta^2 & 0 & 0 & \ldots & 0 & 0 & 0 \\
\beta^2 & \alpha\beta & \beta^2 & \alpha\beta & 0 & 0 & \ldots & 0 & 0 & 0 \\
0 & 0 & \alpha\beta & \beta^2 & \alpha\beta & \beta^2 & \ldots & 0 & 0 & 0 \\
0 & 0 & \beta^2 & \alpha\beta & \beta^2 & \alpha\beta & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & \ldots & \alpha\beta & \beta^2 & \beta \\
0 & 0 & 0 & 0 & 0 & 0 & \ldots & \beta^2 & \alpha\beta & \beta
\end{pmatrix}$$

Extracting a common factor β from each row, and using the kind of transformations explained in Example 15, we obtain the following matrix, which has the same rank:

$$\begin{pmatrix} 1&1&0&0&0&0&\dots&0&0&0\\ 0&1&0&1&0&0&\dots&0&0&0\\ 1&0&1&0&0&0&\dots&0&0&0\\ 0&0&0&1&0&1&\dots&0&0&0\\ 0&0&1&0&1&0&\dots&0&0&0\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\ddots&\vdots&\vdots&\vdots\\ 0&0&0&0&0&0&\dots&0&1&1\\ 0&0&0&0&0&0&\dots&1&0&1 \end{pmatrix}$$

and now it is clear that this matrix has rank $2m+1$, i.e., the cardinal of the orbit under consideration.

Consider now the contribution of cyclic orbits. Each cyclic orbit with two elements, i.e., each contact $i\cdot j \in Q_1 \cap Q_2$, appears in the transfer matrix T_{Γ_1,Γ_2} as the ith and jth row and column with a 1 on the diagonal and the other entries equal to 0, and therefore (as it was the case of isolated nodes) it does not contribute anything either to $\delta(\Gamma_1,\Gamma_2)$ or $\rho(\Gamma_1,\Gamma_2)$. On the other hand, any contact $i\cdot j \in Q_1 \cap Q_2$ does not add anything to $Q_1 \triangle Q_2$, either.

Consider finally a cyclic orbit with cardinal $2m \geq 4$, which corresponds to $2m$ elements in $Q_1 \triangle Q_2$: let us show that it contributes a factor α^{2m} to $\delta(\Gamma_1,\Gamma_2)$ and that it adds up $2m-2$ to $\rho(\Gamma_1,\Gamma_2)$, and we shall be done. Without any loss of generality, we may assume that this orbit is $\{i, i+1, i+2, \dots, i+2m-1\}$ with $i\cdot(i+1), (i+2)\cdot(i+3), \dots, (i+2m-2)\cdot(i+2m-1) \in Q_1$ and $(i+1)\cdot(i+2), \dots, (i+2m-3)\cdot(i+2m-2), i\cdot(i+2m-1) \in Q_2$. As before, a simple computation shows that the minor of T_{Γ_1,Γ_2} corresponding to rows and columns between ith one and the $(i+2m-1)$th one is the following:

$$\begin{pmatrix} \alpha^2&\alpha\beta&0&0&0&0&\dots&\beta^2&\alpha\beta\\ \alpha\beta&\alpha^2&\alpha\beta&\beta^2&0&0&\dots&0&0\\ \beta^2&\alpha\beta&\alpha^2&\alpha\beta&0&0&\dots&0&0\\ 0&0&\alpha\beta&\alpha^2&\alpha\beta&\beta^2&\dots&0&0\\ 0&0&\beta^2&\alpha\beta&\alpha^2&\alpha\beta&\dots&0&0\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\ddots&\vdots&\vdots\\ \alpha\beta&\beta^2&0&0&0&0&\dots&\alpha\beta&\alpha^2 \end{pmatrix}$$

and therefore the product of the entries of its main diagonal is, indeed, α^{2m}.

As far as its contribution to $\rho(\Gamma_1,\Gamma_2)$ goes, if we add up 1 to the entries of its main diagonal, then a set of transformations similar to those carried over in the case of the minor corresponding to a linear orbit produces the following matrix with the same rank:

$$\begin{pmatrix} 1&0&0&0&0&0&\dots&1&0\\ 0&1&0&1&0&0&\dots&0&0\\ 1&0&1&0&0&0&\dots&0&0\\ 0&0&0&1&0&1&\dots&0&0\\ 0&0&1&0&1&0&\dots&0&0\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\ddots&\vdots&\vdots\\ 0&1&0&0&0&0&\dots&0&1 \end{pmatrix}$$

In this matrix, rows with indexes $2, \dots, 2m-2$ are clearly independent, while (the ground field being of characteristic 2) the sum of the odd-indexed rows is zero, as well as the sum of the even-indexed ones. Therefore, the rank of this matrix is indeed $2m-2$. □