Recombination faults in gene assembly in ciliates modeled using multimatroids

Non Peer-reviewed author version

BRIJDER, Robert (2015) Recombination faults in gene assembly in ciliates modeled using multimatroids. In: THEORETICAL COMPUTER SCIENCE, 608, p. 27-35.

DOI: 10.1016/j.tcs.2015.08.038
Handle: http://hdl.handle.net/1942/20703

Recombination Faults in Gene Assembly in Ciliates Modeled Using Multimatroids

Robert Brijder[1]

Hasselt University and Transnational University of Limburg, Belgium

Abstract

We formally model the process of gene assembly in ciliates on the level of individual genes using the notion of multimatroids introduced by Bouchet. Gene assembly involves heavy splicing and recombination, and it turns out that multimatroids form a suitable abstract model that captures essential features of this process. We use this abstract model to study the effect of faulty recombinations during the gene assembly process.

Keywords: gene assembly in ciliates, multimatroids, DNA recombination, circuit partitions, 4-regular graphs

1. Introduction

Gene assembly is an intricate process going on in unicellular organisms called ciliates. Gene assembly involves heavy splicing and recombination operations occurring in a highly parallel fashion [21, 20, 22]. This process has been formally studied on the level of individual genes, see, e.g., [13]. Intramolecular models of gene assembly assume that recombination takes place *within* a molecule (in contrast to intermolecular models [18]). Two well-studied and essentially equivalent models of intramolecular gene assembly are based on (1) particular operations on signed double occurrence strings and (2) local and edge complementation on graphs with loops (or signs) [12, 16], see also [13, 6].

In this paper we formalize gene assembly (on the level of genes) using multimatroids, and in particular we focus on intramolecular gene assembly. Multimatroids have been introduced by Bouchet [4] in an effort to generalize both the theory of circuit partitions in 4-regular multigraphs initiated by Kotzig [17] and to generalize the notions of delta-matroids [3] and isotropic systems [2], the latter of which generalizes properties of pairs of mutually-dual binary matroids. Multimatroids have computationally interesting properties, for example it allows for a particular greedy algorithm. In this paper we show that only very

Email address: robert.brijder@uhasselt.be (Robert Brijder)

[1] R.B. is a postdoctoral fellow of the Research Foundation – Flanders (FWO).

I_1	M_1	I_2	$\overline{M_5}$	I_3	M_4	I_4	$\overline{M_3}$	I_5	$\overline{M_2}$	I_6	M_7	I_7	M_6	I_8

Figure 1: MIC form of a gene

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	

Figure 2: MAC form of a gene

little (multi)matroid theory is necessary to appreciate its power in studying gene assembly. In fact, while helpful, we require in this paper no prior knowledge of matroids or multimatroids. We present an easy (but new) result on multimatroids that generalizes some results in the literature on delta-matroids, isotropic systems, and graphs involving local and loop complementation. Finally we apply this result to show how the set of particular strategies of recombination changes when some erroneous recombination occurs. This is motivated by the fact that various errors and mutations do occur in Nature — in fact mutations are a driving force in evolution.

This paper is organized as follows. In Section 2 we recall the process of gene assembly in ciliates, and in Section 3 we model this process (on the level of genes) using 4-regular multigraphs. In Section 4 we recall the notion of multimatroids are related notions such as a matroid, and we show how multimatroids relate to 4-regular multigraphs and gene assembly in particular. Then, in Section 5 we show how different matroids in a multimatroid that are "close" to each other (their ground sets differ only by a skew pair) are related to each other. We show consequences for recombination faults during gene assembly in Section 6. A discussion (Section 7) concludes this paper.

2. Gene Assembly

In this section we give a concise description of the process of gene assembly on the level of individual genes. We refer to, e.g., [13] for a more gentle and detailed treatment.

During sexual reproduction of unicellular organisms called ciliates, a nucleus, called the micronucleus (MIC for short), is transformed into a structurally and functionally different nucleus called the macronucleus (MAC) in a process called *gene assembly*. On the level of individual genes, each gene is transformed from its MIC form to its MAC form. The MAC form of a gene is able to transcribe, while the MIC form can be seen as a scrambled version of the MAC form that stays dormant.

The MIC form of a gene is a sequence of, possibly inverted (i.e., rotated 180 degrees), *macronuclear destined sequences* (MDSs for short) with *internal eliminated sequences* (IESs for short) in-between, while the MAC form a gene is

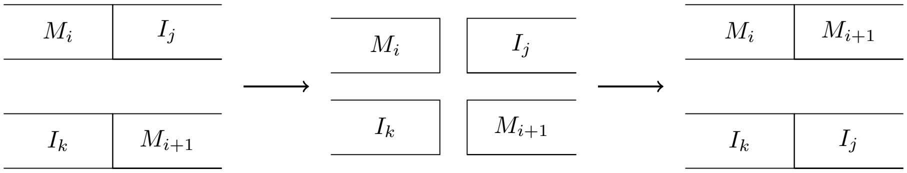

Figure 3: DNA recombination

a consecutive sequence of MDSs that appear in the right order (the MDSs are ordered) and are not inverted[2]. The MIC form and MAC form of an example gene is given in Figures 1 and 2, respectively. The M_i's represent MDSs and the I_i's represent IESs, with bars indicating inversion. Notice that M_2, M_3, and M_5 are inverted in Figure 1. Also, notice that the MDSs in Figure 2 appear in the right order and they are not inverted — as required. This gene is the running example of this paper.

During gene assembly, the MIC form is spliced and recombined to "glue" each two consecutive MDSs; in this way obtaining the MAC form of the gene, see Figure 3. This process, called *recombination*, is one way, i.e., two consecutive MDSs that have been glued together will not be "unglued". During gene assembly, each IES ends up either left or right of the MAC form of the gene or it becomes part of a circular DNA molecule consisting solely of IESs.

3. 4-Regular Multigraphs

A multigraph G is called *Eulerian* when each vertex has even degree. Note that we do not require that G is connected. Also, G is 4-regular when each vertex has degree 4. We allow self-loops, where a self-loop counts for two in the calculation of the degree. A *circuit partition* of a multigraph G is a partition P of the edges such that each $C \in P$ is an (unoriented) circuit of G, where a *circuit* is a closed walk, without orientation, allowing repetitions of vertices but not of edges. The set of vertices of G is denoted by $V(G)$.

We now show that the MIC form and the MAC form of a gene can naturally be captured as circuit partitions within a 4-regular multigraph G. Let us call the positions in the MIC where recombination takes place *recombination points* (also called *pointers* in the literature). Recall from Figure 3 that recombination glues each two MDSs M_i and M_{i+1}. Hence the boundary of the right-hand side of each MDS M_i, except for the last MDS, is a recombination point, say i_a, and the boundary of the left-hand side of each MDS M_i, except for the first MDS, is a recombination point, say $(i-1)_b$. First we represent the MIC form of Figure 1 as a digraph, where each vertex has in-degree and out-degree equal to one, by

[2] Actually, the MDSs in the MAC form of the gene overlap slightly, where the overlapping regions are called *pointers*. This is however irrelevant for this paper, and so, in order to simplify the presentation, we do not discuss pointers here.

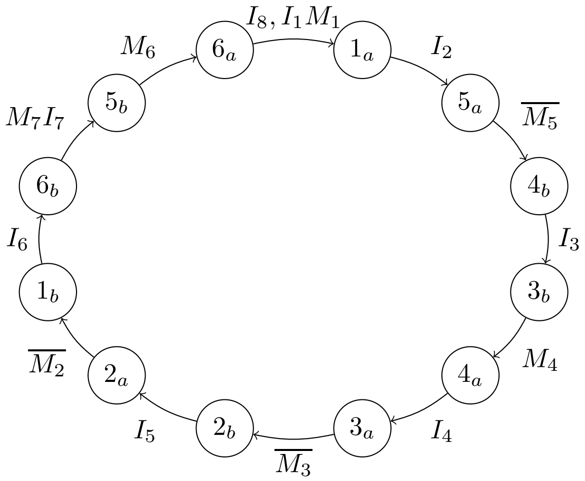

Figure 4: Digraph representing the MIC form of the running example.

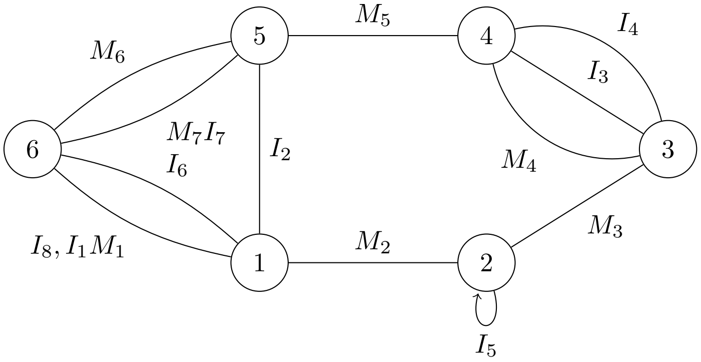

Figure 5: 4-regular graph of the running example.

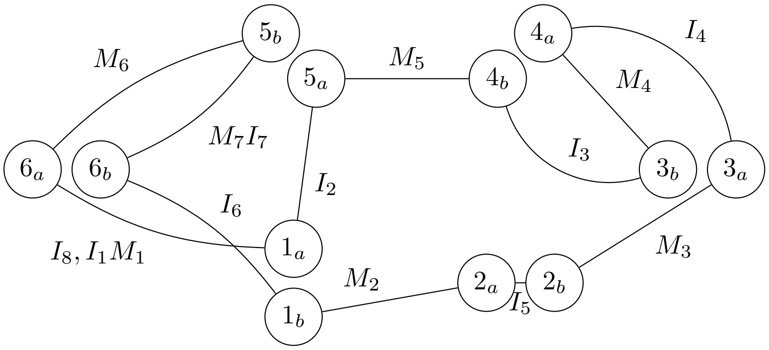

Figure 6: Cycle graph representing the MIC form of the running example.

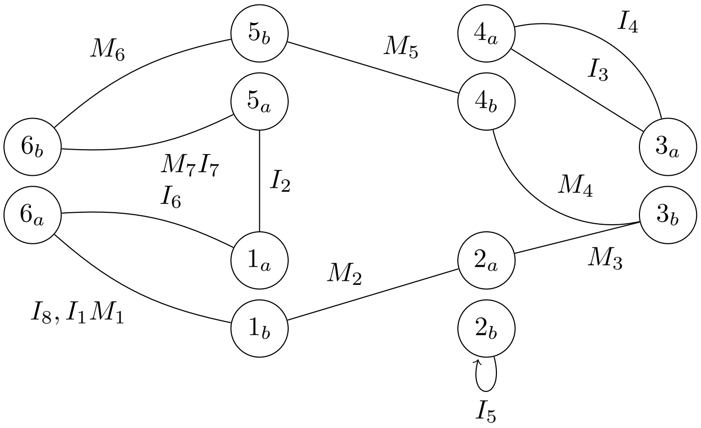

Figure 7: Cycle graph representing the MAC form of the running example.

(1) adding vertices for all recombination points i_a and i_b and (2) adding an arc from each recombination point to the next recombination point, see Figure 4. The arcs are labeled by the MDS/IES segments they represent. The segment I_8 on the right-hand side of the right-most recombination point is merged with the segment $I_1 M_1$ on the left-hand side of the left-most recombination point to obtain a single edge labeled by $I_8, I_1 M_1$. This is done to obtain a cycle (which will lead to a valid 4-regular graph). Now, the 4-regular multigraph of Figure 5 is obtained from Figure 4 by (1) merging the pair of vertices i_a and i_b for all i, (2) forgetting the direction of the edges, and (3) by removing all bars (e.g., M_5 instead of $\overline{M_5}$).

The MIC and MAC forms of the gene belong now to two particular circuit partitions of Figure 5, which are illustrated by the graphs of Figures 6 and 7, respectively. In particular, notice the similarities between Figures 4 and 6. Note that the circuit partition of the MIC form is always an Eulerian circuit. Gene assembly has also been modeled using 4-regular multigraphs in a similar way in [15, 1].

We use terminology from [4]. For a set A, we denote the cardinality of A by $|A|$. We associate to each edge $e = \{v, w\}$ two *half-edges*. One half-edge is incident to v and the other is incident to w. In particular, two half-edges are associated to a self-loop e (which corresponds to the case $v = w$). The half-edges of a graph are mutually distinct, and so the number of half-edges is twice the number of edges. A *local splitter* of G at vertex v of an Eulerian multigraph G is a pair $s_v = \{r_1, r_2\}$, where s_v is a partition of the set of half-edges of G incident to v such that $|r_1|$ and $|r_2|$ are even and nonzero.[3] Thus, if G is a 4-regular graph, then there are three distinct local splitters of G at each vertex v. Indeed, there are precisely three partitions s_v of the set of 4 half-edges $\{h_1, h_2, h_3, h_4\}$ incident to v into two pairs: $\{\{h_1, h_2\}, \{h_3, h_4\}\}$, $\{\{h_1, h_3\}, \{h_2, h_4\}\}$, and $\{\{h_1, h_4\}, \{h_2, h_3\}\}$.

Definition 1. The *detachment* of an Eulerian multigraph G at a local splitter $s_v = \{r_1, r_2\}$ at vertex v, denoted by $G||s_v$, is the multigraph obtained from G by splitting v into two vertices v_1 and v_2 where v_1 is incident to the half-edges of r_1 and v_2 is incident to the half-edges of r_2.

For a vertex v of a 4-regular graph, the three possible detachments at v are illustrated in Figure 8.

Note that if G is Eulerian, then $G||s_v$ is again Eulerian. Also note that the operation of detachment is defined up to isomorphism since the identities of the vertices v_1 and v_2 are not specified. Finally note that detachment commutes for local splitters s_v and s_w at distinct vertices v and w, i.e., $(G||s_v)||s_w = (G||s_w)||s_v$ when v and w are distinct. Hence for a set S of local splitters at mutually distinct vertices, we write $G||S$ to denote $G||s_{v_1} \cdots ||s_{v_n}$ for any order of local splitters of S. As an example, the graph of Figure 6 (Figure 7, resp.)

[3] In [4] local splitters are called nonnull local splitters.

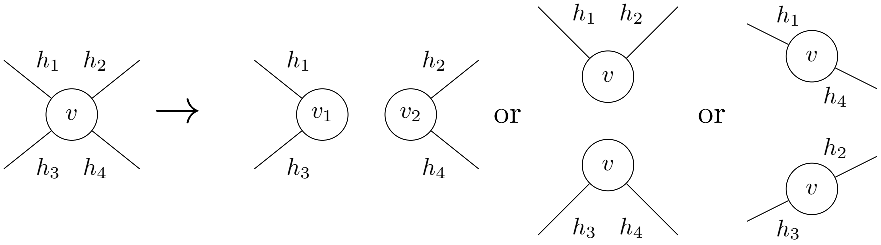

Figure 8: The three possible detachments $G||s_v$ at a degree four vertex v. The three cases correspond to the cases $s_v = \{\{h_1, h_3\}, \{h_2, h_4\}\}$, $s_v = \{\{h_1, h_2\}, \{h_3, h_4\}\}$, and $s_v = \{\{h_1, h_4\}, \{h_2, h_3\}\}$, respectively.

can be obtained from the graph of Figure 5 by detachment at a particular set T_{MIC} (T_{MAC}, resp.) of local splitters for each vertex of G. Let us consider a graph a *cycle graph* when each vertex is of degree 2. Notice that the graphs of Figures 6 and 7 are cycle graphs because detachment has been applied at each vertex of a 4-regular graph. Recall from above that Figures 6 and 7 represent the MIC and MAC form of the gene, respectively. Hence, we may view T_{MIC} and T_{MAC} as the sets of local splitters of G that belong to the MIC and MAC, respectively.

A *carrier* is a tuple (U, Ω) where Ω is a partition of a finite set U, called the *ground set*. Every $\omega \in \Omega$ is called a *skew class*, and a $p \subseteq \omega$ with $|p| = 2$ is called a *skew pair* of ω. A *transversal* (*subtransversal*, resp.) T of Ω is a subset of U such that $|T \cap \omega| = 1$ ($|T \cap \omega| \leq 1$, resp.) for all $\omega \in \Omega$. We denote the set of transversals of Ω by $\mathcal{T}(\Omega)$, and the set of subtransversals of Ω by $\mathcal{S}(\Omega)$.

For a 4-regular multigraph G, we define the carrier $\mathcal{P}_G = (U, \Omega)$, where U the set of local splitters of G and for all $\omega \in \Omega$, $x, y \in \omega$ if and only if x and y are local splitters at a common vertex. Note that for all $\omega \in \Omega$, we have $|\omega| = 3$.

Assume that G belongs to a gene. Let T_{MIC} be the transversal of $\mathcal{P}_G$ belonging to the MIC form of a gene and let T_{MAC} be the transversal of $\mathcal{P}_G$ belonging to its MAC form. Note that $T_{\mathrm{MIC}} \cap T_{\mathrm{MAC}} = \varnothing$. Let $p_v \subseteq T_{\mathrm{MIC}} \cup T_{\mathrm{MAC}}$ be a skew pair belonging to vertex v. Then we notice that $G||(T_{\mathrm{MIC}} \,\Delta\, p_v)$, where Δ denotes symmetric difference, is the cycle graph representing the DNA structure obtained from the MIC form by applying recombination on the recombination points belonging to vertex v. We can continue this process, obtaining a sequence of cycle graphs, until we obtain the cycle graph representing the MAC form of the gene. This process is captured by the notion of a strategy, defined below.

Definition 2. Let G be a 4-regular graph, and let T_{MIC} and T_{MAC} be disjoint transversals of $\mathcal{P}_G$. Let P be a set of mutually disjoint skew pairs such that $\bigcup P = T_{\mathrm{MIC}} \cup T_{\mathrm{MAC}}$. A *strategy* s for G with respect to T_{MIC} and T_{MAC} is an ordered partition of P, i.e., $s = (P_1, \ldots, P_l)$, where the P_i's are nonempty and $\bigcup_i P_i = P$.

Following strategy s, the following cycle graphs occur during gene assembly:

$$G||T_{\mathrm{MIC}}, G||(T_{\mathrm{MIC}} \,\Delta\, P_1), G||(T_{\mathrm{MIC}} \,\Delta\, P_1 \,\Delta\, P_2), \ldots, G||(T_{\mathrm{MIC}} \,\Delta\, P_1 \,\Delta \cdots \Delta\, P_l). \quad (1)$$

I_1	M_2	I_2	M_1	I_3	M_3	I_4

Figure 9: MIC form of a gene

We note that $G||(T_{\mathrm{MIC}} \,\Delta\, P_1 \,\Delta \cdots \Delta\, P_l) = G||T_{\mathrm{MAC}}$. The reason that the P_i's are sets of skew pairs instead of single skew pairs, is due to the fact that more than one recombination operation may take place in parallel. This formalization of a strategy is similar as done in [15] (see also [13, Chapter 14]) in the context of gene assembly.

Definition 3. A *refinement* of a strategy $s = (P_1, \dots, P_l)$ is a strategy $s' = (Q_1, \dots, Q_k)$ such that if $x \in P_i$ and $y \in P_j$ with $i \leq j$, then $x \in Q_{i'}$ and $y \in Q_{j'}$ where both (1) $i' = j'$ implies $i = j$, and (2) $i' \leq j'$. We write in this case $s \leq s'$.

Note that $\leq$ is a partial order.

In the intramolecular model of gene assembly [13], recombination only takes place *within* molecules. In this model, the strategies that are both intramolecular and maximal with respect to the $\leq$ relation are of particular interest. This notion is captured by the following definition. Let us denote the number of connected components of a multigraph G by $k(G)$.

Definition 4. Let $s = (P_1, \dots, P_l)$ be a strategy for G with respect to disjoint transversals T_{MIC} and T_{MAC} of $\mathcal{P}_G$. Then s is called *intramolecular* if for all $i \in \{0, \dots, l\}$ and for all $p \in \cup_{j \in \{i+1, \dots, l\}} P_j$, $k(G_i) \leq k(G_i||p)$ where $G_i = G||(T_{\mathrm{MIC}} \,\Delta\, P_1 \,\Delta \cdots \Delta\, P_i)$.

Moreover we say that s is *maximal intramolecular* if s is intramolecular and for all intramolecular strategies $s \leq s'$, we have $s = s'$.

It is shown in [14] (see also [13]) using strings instead of graphs, that for each 4-regular multigraph and disjoint transversals T_{MIC} and T_{MAC} there is a intramolecular strategy where each of the P_i's are of cardinality at most 2.

This number 2 cannot be reduced in general (i.e., there does not always exist an intramolecular strategy consisting of only singletons), as the following example illustrates.

Example 5. Consider the MIC form of a gene of Figure 9. Note that two recombinations are necessary to transform this gene to its MAC form: one recombination glues M_1 to M_2 and another recombination glues M_2 to M_3. It is easy to see that the only maximal intramolecular strategy for this gene is by applying both recombinations in parallel. Indeed, applying only one of the two recombination operations would split the molecule in two, while the end result, i.e., the MAC form of the gene, consists of one molecule. Therefore, there exists exactly one maximal intramolecular strategy $s = (P_1)$ for the corresponding 4-regular multigraph G, where P_1 contains the two local splitters corresponding to the two recombinations. We remark that s corresponds to applying one so-called double loop, alternating direct-repeat excision-reinsertion operation (dlad for short), see, e.g., [13] for its definition. Moreover, any strategy s' of G of the form (P'_1, P'_2) (where P'_1 and P'_2 are therefore singletons) is not an intramolecular strategy. □

Of course, $\mathcal{P}_G$ alone cannot tell which strategies are maximal intramolecular and which are not. In fact, almost all information is lost when considering $\mathcal{P}_G$ alone. We now extend the tuple $\mathcal{P}_G$ to a triple $\mathcal{Q}_G$ that retains more information regarding properties of the various possible strategies.

Definition 6. Let G be a 4-regular multigraph. We define $\mathcal{Q}_G$ to be the triple $(U, \Omega, \mathcal{C})$ where $\mathcal{P}_G = (U, \Omega)$ and $\mathcal{C} \subseteq \mathcal{S}(\Omega)$ such that for $C \in \mathcal{S}(\Omega)$ we have $C \in \mathcal{C}$ if and only if C is minimal (with respect to inclusion) such that $G||C$ has a larger number of connected components than G.

Example 7. We continue the running example. Consider $\mathcal{Q}_G = (U, \Omega, \mathcal{C})$ with G as in Figure 5. We denote, for all vertices v of G, by v_{MIC} the transition at vertex v that is taken in the MIC form of the gene, cf. Figure 6, by v_{MAC} the transition at vertex v that is taken in the MAC form of the gene, cf. Figure 7, and by v_{ERR} the third, "faulty", transition at v (distinct from both v_{MIC} and v_{MAC}). Then

$$\begin{aligned}\mathcal{C} = \{ \quad & \{2_{\mathrm{MAC}}\}, \{3_{\mathrm{MAC}}, 4_{\mathrm{MAC}}\}, \{5_{\mathrm{MIC}}, 6_{\mathrm{MAC}}\}, \{1_{\mathrm{MAC}}, 5_{\mathrm{MAC}}, 6_{\mathrm{MIC}}\}, \\ & \{1_{\mathrm{ERR}}, 5_{\mathrm{MIC}}\}, \{1_{\mathrm{ERR}}, 6_{\mathrm{MAC}}\}, \{3_{\mathrm{MIC}}, 4_{\mathrm{ERR}}\}, \{3_{\mathrm{ERR}}, 4_{\mathrm{MIC}}\}, \\ & \{1_{\mathrm{MIC}}, 5_{\mathrm{ERR}}, 6_{\mathrm{MIC}}\}, \{1_{\mathrm{MAC}}, 5_{\mathrm{ERR}}, 6_{\mathrm{ERR}}\}, \{1_{\mathrm{MIC}}, 5_{\mathrm{MAC}}, 6_{\mathrm{ERR}}\}\}.\end{aligned}$$

Note that $\{3_{\mathrm{MAC}}, 4_{\mathrm{MAC}}\} \in \mathcal{C}$ and $\{2_{\mathrm{MAC}}\} \in \mathcal{C}$ are easily seen from Figure 7. □

While G cannot be completely reconstructed (not even up to isomorphism) from $\mathcal{Q}_G$, we will show in the next sections that $\mathcal{Q}_G$ retains key properties of the various possible strategies.

4. Matroids and Multimatroids

In this section we make abstract essential properties of the effect of splitting vertices using detachments [4]. It turns out that these essential properties are elegantly captured using matroids and multimatroids. Although helpful, we do not require in this paper prior knowledge of matroids or multimatroids. For a more elaborate exposition of the notions and terminology concerning matroids, we refer to the monographs [19, 23].

4.1. Matroids

Matroids can be defined in various different ways, such as in terms of rank, bases, circuits, independent sets, etc. We define matroids here in terms of circuits. The power set of a set X is denoted by 2^X.

Definition 8 ([4]). A *matroid* M (described by its circuits) is a tuple $(E, \mathcal{C})$ where E is a finite set called the *ground set* of M and $\mathcal{C} \subseteq 2^E$ such that

1. $\varnothing \notin \mathcal{C}$,
2. if $C_1, C_2 \in \mathcal{C}$ with $C_1 \subseteq C_2$, then $C_1 = C_2$, and
3. if $C_1, C_2 \in \mathcal{C}$ are distinct and $x \in C_1 \cap C_2$, then there is a $C \in \mathcal{C}$ with $C \subseteq (C_1 \cup C_2) \setminus \{x\}$.

The elements of $\mathcal{C}$ are called the *circuits* of M. We warn the reader that the notion of circuit used here (and in the context of multimatroids in the next subsection) is different from the notion of circuit in the context of 4-regular multigraphs. We denote the ground set of M by $E(M)$ and the set of circuits of M by $\mathcal{C}(M)$. For any $X \subseteq E(M)$, the *restriction* of M to X, denoted by $M[X]$, is the matroid $(X, \mathcal{C} \cap 2^X)$. For $x \in E$ we write $M \setminus x$ for $M[E(M) \setminus \{x\}]$.

A set $B \subseteq E(M)$ is called a *basis* of M if B does not contain any circuit of M and is maximal (with respect to inclusion) with this property. It turns out that the bases of a matroid have a common cardinality called the *rank* of M, denoted by $r(M)$. The *nullity* of M, denoted by $n(M)$, is $|E(M)| - r(M)$.

A $x \in E$ is called a *coloop* of M if x is not contained in any circuit. If $x \notin E(M)$, then $M \oplus x$ denotes adding x as a coloop to M. Note that $\mathcal{C}(M \oplus x) = \mathcal{C}(M)$ and $E(M \oplus x) = E(M) \cup \{x\}$.

4.2. Multimatroids

We recall now the notion of a multimatroid and related notions from [4]. Like matroids, multimatroids can be defined in terms of rank, circuits, independent sets, etc. We define multimatroids here in terms of circuits.

Definition 9 ([4]). A *multimatroid* Q (described by its circuits) is a triple $(U, \Omega, \mathcal{C})$, where (U, Ω) is a carrier and $\mathcal{C} \subseteq \mathcal{S}(\Omega)$ such that:

1. for each $T \in \mathcal{T}(\Omega)$, $(T, \mathcal{C} \cap 2^T)$ is a matroid (described by its circuits) and
2. if $C_1, C_2 \in \mathcal{C}$, then $C_1 \cup C_2$ does not include precisely one skew pair.

The elements of $\mathcal{C}$ are called the *circuits* of Q. For any $X \subseteq U$, the *restriction* of Q to X, denoted by $Q[X]$, is the multimatroid $(X, \Omega', \mathcal{C} \cap 2^X)$ with $\Omega' = \{\omega \cap X \mid \omega \cap X \neq \varnothing, \omega \in \Omega\}$. If X is a subtransversal, then we identify $Q[X]$ with the matroid $(X, \mathcal{C} \cap 2^X)$ since $\Omega' = \{\{u\} \mid u \in X\}$ captures no additional information.

A *projection* of carrier (U, Ω) is a surjective function $\pi : U \to V$ such that $\pi(x) = \pi(y)$ if and only if x and y are in the same skew class $\omega \in \Omega$. Thus each skew class is assigned by π to a unique element of V. If $\pi : U \to V$ is a projection, then we also say that Q is *indexed* on V by π. As usual, we let for $X \subseteq U$, $\pi(X) = \{\pi(x) \mid x \in X\}$. Let S be a subtransversal of Ω. The isomorphic image of the matroid $Q[S]$ induced by π (i.e., the renaming of the ground set according to π) is called the *projection* of $Q[S]$ by π and is denoted by $\pi(Q[S])$. More explicitly, if $Q[S] = (S, \mathcal{C})$, then $\pi(Q[S]) = (\pi(S), \{\pi(C) \mid C \in \mathcal{C}\})$.

4.3. Back to 4-regular multigraphs and gene assembly

It has been shown in [4] that $\mathcal{Q}_G$ defined at the end of Section 3 is a multimatroid.

Theorem 10 ([4]). *Let G be a 4-regular multigraph. Then $\mathcal{Q}_G$ is a multimatroid.*

We say that $\mathcal{Q}_G$ is the *multimatroid of G*.

Note again that the circuits of $\mathcal{Q}_G$ are not to be confused with the circuits of G.

Example 11. We continue the running example. Let T_{MAC} (T_{MIC}, resp.) be the transversal of $\mathcal{Q}_G$ containing v_{MAC} (v_{MIC}, resp.) for all vertices v of G. Then $\mathcal{Q}_G[T_{\mathrm{MAC}}]$ is the matroid $(T_{\mathrm{MAC}}, \{\{3_{\mathrm{MAC}}, 4_{\mathrm{MAC}}\}, \{2_{\mathrm{MAC}}\}\})$. The family of bases of $\mathcal{Q}_G[T_{\mathrm{MAC}}]$ is

$$\{\{1_{\mathrm{MAC}}, 4_{\mathrm{MAC}}, 5_{\mathrm{MAC}}, 6_{\mathrm{MAC}}\}, \{1_{\mathrm{MAC}}, 3_{\mathrm{MAC}}, 5_{\mathrm{MAC}}, 6_{\mathrm{MAC}}\}\}.$$

□

Remark 12. *We remark that $\mathcal{Q}_G$ turns out to fulfill a special property called tightness [5], which ensures by [9, Theorem 13] that $\mathcal{Q}_G$ can be uniquely determined by $\mathcal{Q}_G[T_{\mathrm{MIC}} \cup T_{\mathrm{MAC}}]$. In the running example, the family of circuits of $\mathcal{Q}_G[T_{\mathrm{MIC}} \cup T_{\mathrm{MAC}}]$ is*

$$\{\{2_{\mathrm{MAC}}\}, \{3_{\mathrm{MAC}}, 4_{\mathrm{MAC}}\}, \{5_{\mathrm{MIC}}, 6_{\mathrm{MAC}}\}, \{1_{\mathrm{MAC}}, 5_{\mathrm{MAC}}, 6_{\mathrm{MIC}}\}\}.$$

It turns out that multimatroids elegantly capture many interesting properties of circuit partitions in 4-regular multigraphs [4]. For example, a *basis* of a multimatroid of Q is a $B \in \mathcal{S}(\Omega)$ that is maximal (with respect to inclusion) such that it does not contain any circuit of Q. It turns out that the family of bases of $\mathcal{Q}_G$ correspond precisely to circuit partitions that are the Eulerian systems of G (an Eulerian system is a set of Eulerian circuits, one for each connected component).

It is shown in [4] that for all $S \in \mathcal{S}(\Omega)$, $n(\mathcal{Q}_G[S]) = k(G||S) - k(G)$, where we recall that $k(\cdot)$ denotes the number of connected components of a multigraph. While by definition the circuits of $\mathcal{Q}_G$ are the minimal sets such that $k(G||C) - k(G) > 0$, it is interesting to observe that this nullity property shows that the circuits of $\mathcal{Q}_G$ actually determine the value $k(G||S) - k(G)$ for all $S \in \mathcal{S}(\Omega)$! This is a prime example of the power of multimatroids when studying formal properties of gene assembly.

Using this nullity result we can formulate the notion of an intramolecular strategy completely in terms of the multimatroid $\mathcal{Q}_G$ as follows.

Corollary 13. *Let G be a 4-regular multigraph and let $s = (P_1, \ldots, P_l)$ be a strategy for G with respect to some disjoint transversals T_{MIC} and T_{MAC}. Then s is intramolecular if and only if for all $i \in \{0, \ldots, l\}$ and for all $p \in \cup_{j \in \{i+1,\ldots,l\}} P_j$, $n(\mathcal{Q}_G[S_i]) \leq n(\mathcal{Q}_G[S_i \,\Delta\, p])$ where $S_i = T_{\mathrm{MIC}} \,\Delta\, P_1 \,\Delta\, P_2 \cdots \Delta\, P_i$.*

If the 4-regular multigraph G corresponds to a gene, then $n(\mathcal{Q}_G[T_{\mathrm{MIC}}]) = 0$ as the MIC form of the gene is one molecule (and thus forms an Eulerian circuit in G) and $n(\mathcal{Q}_G[T_{\mathrm{MAC}}])$ is one less than the number of molecules obtained after gene assembly, one of which contains the MAC form of the gene.

Example 14. In the running example, we have $n(\mathcal{Q}_G[T_{\mathrm{MAC}}]) = 2$. This follows from Figure 7 as it contains three connected components and it also follows from Example 11 where it is shown that the bases of $\mathcal{Q}_G[T_{\mathrm{MAC}}]$ are all cardinality 4. Hence $n(\mathcal{Q}_G[T_{\mathrm{MAC}}]) = |V(G)| - 4 = 2$. □

5. General Multimatroid Result

With multimatroids as a suitable abstraction of key properties of 4-regular multigraphs in place, we formulate in this section the main technical result of this paper. First we provide a characterization of multimatroids.

Lemma 15. *Let $Q = (U, \Omega, \mathcal{C})$ with (U, Ω) a carrier and $\mathcal{C} \subseteq \mathcal{S}(\Omega)$. Denote for $S \in \mathcal{S}(\Omega)$, $Q[S] = (S, \mathcal{C} \cap 2^S)$. Then Q is a multimatroid if and only if*

1. *for each $T \in \mathcal{T}(\Omega)$, $Q[T]$ is a matroid and*
2. *for all $S \in \mathcal{S}(\Omega)$ and $\omega \in \Omega$ with $\omega \cap S = \varnothing$, there is at most one $x \in \omega$ with $\mathcal{C}(Q[S \cup \{x\}]) \neq \mathcal{C}(Q[S])$.*

Proof. First assume that Q is a multimatroid. Let $S \in \mathcal{S}(\Omega)$ and $\omega \in \Omega$ with $\omega \cap S = \varnothing$. We have $\mathcal{C}(Q[S]) \subseteq \mathcal{C}(Q[S \cup \{x\}])$ for all $x \in \omega$. If there are distinct $x_1, x_2 \in \omega$ with $\mathcal{C}(Q[S]) \subsetneq \mathcal{C}(Q[S \cup \{x_i\}])$ for all $i \in \{1, 2\}$, then there are circuits C_1 and C_2 such that $C_1 \cup C_2$ contains the skew pair $\{x_1, x_2\}$ but no other skew pair. Thus, the second condition of the definition of a multimatroid is violated. Hence there is at most one $x \in \omega$ with $\mathcal{C}(Q[S \cup \{x\}]) \neq \mathcal{C}(Q[S])$.

Conversely, assume that the right-hand side of the equivalence holds. It suffices to show that the second condition of the definition of a multimatroid holds. Let $C_1, C_2 \in \mathcal{C}$ such that $C_1 \cup C_2$ contains precisely one skew pair p. Now the second condition of the right-hand side of the equivalence is violated with $S = (C_1 \cup C_2) \setminus p$ and $\omega \in \Omega$ the unique skew class of Ω with $p \subseteq \omega$. □

The important conceptual difference between Lemma 15 and the definition of a multimatroid is that the former relates the entire families of circuits of matroids $Q[S \cup \{x\}]$ and $Q[S]$, while the latter relates individual circuits.

By Lemma 15, $Q[S \cup \{x\}]$ for all $x \in \omega$ are all mutually isomorphic except for at most one. The next result essentially formalizes this observation. It is the main technical result of this paper.

Theorem 16. *Let Q be a multimatroid with carrier (U, Ω), and let $p : U \to V$ be a projection of Q. Let $S \in \mathcal{S}(\Omega)$, and let z be a skew pair of Ω with $z \cap S \neq \varnothing$.*

1. *If $r(Q[S]) \leq r(Q[S \,\Delta\, z])$, then $p(Q[S]) \setminus v \oplus v = p(Q[S \,\Delta\, z])$ where $p(z) = \{v\}$.*

2. *If $r(Q[S]) = r(Q[S \,\Delta\, z])$, then we have $\mathcal{C}(Q[S]) = \mathcal{C}(Q[S \,\Delta\, z])$, $p(Q[S]) = p(Q[S \,\Delta\, z])$, and for all $C \in \mathcal{C}(Q[S])$, $C \cap z = \varnothing$.*

Proof. If $r(Q[S]) \leq r(Q[S \,\Delta\, z])$, then by Lemma 15 $\mathcal{C}(Q[S \,\Delta\, z]) = \mathcal{C}(Q[S \backslash z])$. Hence $\mathcal{C}(Q[S \,\Delta\, z]) = \{C \in \mathcal{C}(Q[S]) \mid C \cap z = \varnothing\}$ and therefore $p(Q[S]) \setminus v \oplus v = p(Q[S \,\Delta\, z])$.

If $r(Q[S]) = r(Q[S \,\Delta\, z])$, then by the first statement, $\mathcal{C}(Q[S]) = \mathcal{C}(Q[S \backslash z]) = \mathcal{C}(Q[S \,\Delta\, z])$. Since the sets of circuits of $Q[S]$ and $Q[S \,\Delta\, z]$ are identical and since p maps S and $S \,\Delta\, z$ to the same set we have $p(Q[S]) = p(Q[S \,\Delta\, z])$. □

We remark that it is an easy consequence of the definition of a multimatroid in terms of rank [4] (we do not recall this definition here) that $r(Q[S])$ and $r(Q[S \,\Delta\, z])$ differ by at most 1.

We stress that the proof of Theorem 16 is very short and with hindsight straightforward only because its formulation is in terms of multimatroids described by its circuits. In less general notions than multimatroids (or even multimatroids described, say, by their bases) the result is less obvious. In fact Theorem 16 is surprising and not obvious in various special cases that appear in the literature, and have significant longer proofs there. For example, it has been shown in [4] that isotropic systems (we not do recall its definition here) may be viewed as a special class of multimatroids. Theorem 16 generalizes Theorem 9.4 of [2] for isotropic systems. As another example, it has been shown in [9] that vf-safe delta-matroids (again, we not do recall its definition here) may also be viewed as a special class of multimatroids. Theorem 16 generalizes Theorem 5.5 of [8] for vf-safe delta-matroids. Theorem 5.5 of [8] is in turn a generalization of (an essential part of) Theorem 25 of [10] concerning the graph operations of local complementation and loop complementation, see [7, 8] for the correspondence between these graph operations and vf-safe delta-matroids.

6. Back to Gene Assembly: Faults in Recombination

We now describe the consequences of Theorem 16 for gene assembly. Let G be the 4-regular multigraph where transversals T_{MIC} and T_{MAC} represent the MIC and MAC forms of a gene, respectively.

We noticed already that the matroid $\mathcal{Q}_G[T_{\mathrm{MAC}}]$ captures information about the MAC gene as $n(\mathcal{Q}_G[T_{\mathrm{MAC}}]$ describes the number of molecules obtained after gene assembly. The next result shows that $\mathcal{Q}_G[T_{\mathrm{MAC}}]$ also captures information about how the MAC gene is obtained during gene assembly. In other words, it also captures some information about the possible strategies.

Proposition 17. *Let G be a 4-regular multigraph and T_{MIC} and T_{MAC} disjoint transversals with $n(\mathcal{Q}_G[T_{\mathrm{MIC}}]) = 0$. Then B is a basis of $\mathcal{Q}_G[T_{\mathrm{MAC}}]$ if and only if there is a strategy $s_B = (P_0, \ldots, P_l)$ where P_0 contains those skew pairs in $T_{\mathrm{MIC}} \cup T_{\mathrm{MAC}}$ that intersect with B and the other P_i's are singletons that do not intersect with B, such that $n(\mathcal{Q}_G[T_{\mathrm{MIC}} \,\Delta\, P_0 \,\Delta \cdots P_i]) = i$ for all $i \in \{0, \ldots, l\}$.*

PROOF. First assume that B is a basis of $\mathcal{Q}_G[T_{\mathrm{MAC}}]$. Therefore, $\mathcal{Q}_G[B]$ does not contain any circuits. By Lemma 15, for every $\omega \in \Omega$ with $\omega \cap B = \varnothing$, there is at most one $x \in \omega$ with $\mathcal{C}(Q[B \cup \{x\}]) \neq \mathcal{C}(Q[B])$. Since B is a basis, $x \in T_{\mathrm{MAC}}$. Consequently, $\mathcal{Q}_G[T_{\mathrm{MIC}} \,\Delta\, P_0]$ does not have any circuits, and so $n(\mathcal{Q}_G[T_{\mathrm{MIC}} \,\Delta\, P_0]) = 0$. Since by definition of nullity, $n(\mathcal{Q}_G[T_{\mathrm{MAC}}]) = |T_{\mathrm{MAC}}| - r(\mathcal{Q}_G[T_{\mathrm{MAC}}]) = |T_{\mathrm{MAC}}| - |B|$, we have $l = n(\mathcal{Q}_G[T_{\mathrm{MAC}}])$. By the comment below Theorem 16, $n(\mathcal{Q}_G[T_{\mathrm{MIC}} \,\Delta\, P_0 \,\Delta \cdots P_i])$ and $n(\mathcal{Q}_G[T_{\mathrm{MIC}} \,\Delta\, P_0 \,\Delta \cdots P_{i-1}])$ differ by at most 1 for all $i \in \{1, \ldots, l\}$. Hence $n(\mathcal{Q}_G[T_{\mathrm{MIC}} \,\Delta\, P_0 \,\Delta \cdots P_i]) = i$ for all $i \in \{0, \ldots, l\}$.

Conversely, if $n(\mathcal{Q}_G[T_{\mathrm{MIC}} \,\Delta\, P_0]) = 0$, then B does not contain any circuits and if $n(\mathcal{Q}_G[T_{\mathrm{MAC}}]) = l$, then B is maximal with this property. Hence B is a basis of $\mathcal{Q}_G[T_{\mathrm{MAC}}]$. □

Proposition 17 shows that one can divide a strategy into two parts. For each basis B of the matroid $\mathcal{Q}_G[T_{\mathrm{MAC}}]$, one can first recombine on the vertices determined by B to transform the MIC form of a gene to an intermediate form with only one molecule (since $n(\mathcal{Q}_G[T_{\mathrm{MIC}} \,\Delta\, P_0]) = 0$), and then recombine on the remaining vertices where in each step the number of molecules is increased (due to the splitting of a molecule). Hence the bases of $\mathcal{Q}_G[T_{\mathrm{MAC}}]$ characterize the single-molecule intermediate forms that are "closest" to the MAC form of the gene.

Example 18. In the running example, we recall from Example 7 that $B = \{1_{\mathrm{MAC}}, 4_{\mathrm{MAC}}, 5_{\mathrm{MAC}}, 6_{\mathrm{MAC}}\}$ is a basis of $\mathcal{Q}_G[T_{\mathrm{MAC}}]$. Hence by Proposition 17, $n(\mathcal{Q}_G[T_0]) = 0$ with $T_0 = B \cup \{2_{\mathrm{MIC}}, 3_{\mathrm{MIC}}\}$, $n(\mathcal{Q}_G[T_1]) = 1$ with $T_1 = B \cup \{2_{\mathrm{MIC}}, 3_{\mathrm{MAC}}\}$, and $n(\mathcal{Q}_G[T_2]) = 2$ with $T_2 = B \cup \{2_{\mathrm{MAC}}, 3_{\mathrm{MAC}}\} = T_{\mathrm{MAC}}$. □

Note that T_{MAC} can only be obtained from T_{MIC} if for each vertex v of G the "right" skew pair $p_v \subseteq T_{\mathrm{MIC}} \cup T_{\mathrm{MAC}}$ is chosen during gene assembly. Assume that some fault occurs during recombination: by coincidence (perhaps due to some spontaneous mutation) for some vertex v of G the wrong skew pair $p_v \subseteq T_{\mathrm{MIC}} \cup T_{\mathrm{ERR}}$ is chosen. Then we obtain instead $T_{\mathrm{MAC}} \,\Delta\, p_v$ as the end result. Theorem 16 now implies that either (1) $p(\mathcal{Q}_G[T_{\mathrm{MAC}} \,\Delta\, p_v]) = p(\mathcal{Q}_G[T_{\mathrm{MAC}}])$ or (2) $n(\mathcal{Q}_G[T_{\mathrm{MAC}} \,\Delta\, p_v])$ and $n(\mathcal{Q}_G[T_{\mathrm{MAC}}])$ differ by 1, where $p : U \to V(G)$ is the function that maps each local splitter s_v at v to v. In case (1) the strategies as given in Proposition 17 coincide for the "original" T_{MAC} and the "faulty" $T_{\mathrm{MAC}} \,\Delta\, p_v$. However, in case (2) if $n(\mathcal{Q}_G[T_{\mathrm{MAC}} \,\Delta\, p_v]) = n(\mathcal{Q}_G[T_{\mathrm{MAC}}]) - 1$, then for every basis B of $\mathcal{Q}_G[T_{\mathrm{MAC}} \,\Delta\, p_v]$, $B \setminus \{v\}$ is a basis of $\mathcal{Q}_G[T_{\mathrm{MAC}}]$ and this will be reflected in the strategies possible of the form described in Proposition 17. A similar conclusion can be made if $n(\mathcal{Q}_G[T_{\mathrm{MAC}} \,\Delta\, p_v]) - 1 = n(\mathcal{Q}_G[T_{\mathrm{MAC}}])$.

Example 19. In the running example, assume that the faulty transition 5_{ERR} is taken at vertex 5. Then we obtain eventually, $\mathcal{Q}_G[T_{\mathrm{MAC}} \,\Delta \{5_{\mathrm{MAC}}, 5_{\mathrm{ERR}}\}] = \mathcal{Q}_G[\{1_{\mathrm{MAC}}, 2_{\mathrm{MAC}}, 3_{\mathrm{MAC}}, 4_{\mathrm{MAC}}, 5_{\mathrm{ERR}}, 6_{\mathrm{MAC}}\}]$ for which its family of circuits is equal to that of $\mathcal{Q}_G[T_{\mathrm{MAC}}]$. This can be verified straightforwardly from the family of circuits of $\mathcal{Q}_G$ given in Example 7 or by inspecting Figure 5. Hence the two matroids are isomorphic.

If, on the other hand, we assume that the faulty transition 1_{ERR} is taken at vertex 1, then the family of circuits of $\mathcal{Q}_G[T]$ with $T = T_{\mathrm{MAC}}\,\Delta\{1_{\mathrm{MAC}}, 1_{\mathrm{ERR}}\} = \{1_{\mathrm{ERR}}, 2_{\mathrm{MAC}}, 3_{\mathrm{MAC}}, 4_{\mathrm{MAC}}, 5_{\mathrm{MAC}}, 6_{\mathrm{MAC}}\}$ is *not* equal to that of $\mathcal{Q}_G[T_{\mathrm{MAC}}]$. Indeed, the family of circuits of $\mathcal{Q}_G[T]$ is equal to

$$\{\{3_{\mathrm{MAC}}, 4_{\mathrm{MAC}}\}, \{2_{\mathrm{MAC}}\}, \{1_{\mathrm{ERR}}, 6_{\mathrm{MAC}}\}\},$$

and we notice that $\{1_{\mathrm{ERR}}, 6_{\mathrm{MAC}}\}$ is a circuit of $\mathcal{Q}_G[T]$ but not a circuit of $\mathcal{Q}_G[T_{\mathrm{MAC}}]$. Now, the family of bases of $\mathcal{Q}_G[T]$ is equal to

$$\{\{4_{\mathrm{MAC}}, 5_{\mathrm{MAC}}, 6_{\mathrm{MAC}}\}, \{1_{\mathrm{ERR}}, 4_{\mathrm{MAC}}, 5_{\mathrm{MAC}}\}, \{3_{\mathrm{MAC}}, 5_{\mathrm{MAC}}, 6_{\mathrm{MAC}}\},$$
$$\{1_{\mathrm{ERR}}, 3_{\mathrm{MAC}}, 5_{\mathrm{MAC}}\}\}$$

as so we have $n(\mathcal{Q}_G[T]) = |V(G)| - 3 = 3$. □

7. Discussion

We have shown that multimatroids form an elegant abstract model to study the formal properties of gene assembly in ciliates. Moreover, the main technical result (Theorem 16) shows that matroids within a multimatroid for which their ground sets differ only by a skew pair are very closely related. They are either isomorphic or one can be obtained up to isomorphism from the other by deletion and adding a coloop. Finally we showed its consequence for strategies of gene assembly and in particular we showed the differences in strategies in case of one recombination fault.

It turns out that MIC forms of distinct genes may be interleaved in the micronucleus [11]. It would be interesting to investigate the consequences of this observation from a theoretical and computational point of view, and in particular to investigate its consequences in the abstract setting of multimatroids.

Acknowledgements

We thank Lorenzo Traldi for his useful comments on an earlier version of Theorem 16. We also thank Nataša Jonoska for pointing out and explaining the interleavement of distinct MIC genes. Finally, we thank the anonymous referees for their helpful comments on the paper.

References

[1] A. Angeleska, N. Jonoska, and M. Saito. DNA recombination through assembly graphs. *Discrete Applied Mathematics*, 157(14):3020–3037, 2009.

[2] A. Bouchet. Isotropic systems. *European Journal of Combinatorics*, 8:231–244, 1987.

[3] A. Bouchet. Representability of Δ-matroids. In *Proceedings of the 6th Hungarian Colloquium of Combinatorics, Colloquia Mathematica Societatis János Bolyai*, volume 52, pages 167–182. North-Holland, 1987.

[4] A. Bouchet. Multimatroids I. Coverings by independent sets. *SIAM Journal on Discrete Mathematics*, 10(4):626–646, 1997.

[5] A. Bouchet. Multimatroids III. Tightness and fundamental graphs. *European Journal of Combinatorics*, 22(5):657–677, 2001.

[6] R. Brijder, M. Daley, T. Harju, N. Jonoska, I. Petre, and G. Rozenberg. Computational nature of gene assembly in ciliates. In G. Rozenberg, T.H.W. Bäck, and J.N. Kok, editors, *Handbook of Natural Computing*, pages 1233–1280. Springer, 2012.

[7] R. Brijder and H.J. Hoogeboom. The group structure of pivot and loop complementation on graphs and set systems. *European Journal of Combinatorics*, 32:1353–1367, 2011.

[8] R. Brijder and H.J. Hoogeboom. Nullity and loop complementation for delta-matroids. *SIAM Journal on Discrete Mathematics*, 27:492–506, 2013.

[9] R. Brijder and H.J. Hoogeboom. Interlace polynomials for multimatroids and delta-matroids. *European Journal of Combinatorics*, 40:142–167, 2014.

[10] R. Brijder, H.J. Hoogeboom, and L. Traldi. The adjacency matroid of a graph. *The Electronic Journal of Combinatorics*, 20:P27, 2013.

[11] X. Chen, J.R. Bracht, A.D. Goldman, E. Dolzhenko, D.M. Clay, E.C. Swart, D.H. Perlman, T.G. Doak, A. Stuart, C.T. Amemiya, R.P. Sebra, and L.F. Landweber. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell*, 158:1187–1198, 2014.

[12] A. Ehrenfeucht, T. Harju, I. Petre, D.M. Prescott, and G. Rozenberg. Formal systems for gene assembly in ciliates. *Theoretical Computer Science*, 292:199–219, 2003.

[13] A. Ehrenfeucht, T. Harju, I. Petre, D.M. Prescott, and G. Rozenberg. *Computation in Living Cells – Gene Assembly in Ciliates*. Springer Verlag, 2004.

[14] A. Ehrenfeucht, T. Harju, I. Petre, and G. Rozenberg. Characterizing the micronuclear gene patterns in ciliates. *Theory of Computing Systems*, 35:501–519, 2002.

[15] A. Ehrenfeucht, T. Harju, and G. Rozenberg. Gene assembly through cyclic graph decomposition. *Theoretical Computer Science*, 281:325 – 349, 2002.

[16] A. Ehrenfeucht, I. Petre, D.M. Prescott, and G. Rozenberg. String and graph reduction systems for gene assembly in ciliates. *Mathematical Structures in Computer Science*, 12:113–134, 2002.

[17] A. Kotzig. Eulerian lines in finite 4-valent graphs and their transformations. In *Theory of graphs, Proceedings of the Colloquium, Tihany, Hungary, 1966*, pages 219–230. Academic Press, New York, 1968.

[18] L.F. Landweber and L. Kari. The evolution of cellular computing: Nature's solution to a computational problem. *Biosystems*, 52:3–13, 1999.

[19] J.G. Oxley. *Matroid theory, Second Edition.* Oxford University Press, 2011.

[20] D.M. Prescott. Genome gymnastics: Unique modes of DNA evolution and processing in ciliates. *Nature Reviews*, 1:191–199, 2000.

[21] D.M. Prescott and M. DuBois. Internal eliminated segments (IESs) of oxytrichidae. *Journal of Eukaryotic Microbiology*, 43:432–441, 1996.

[22] D.M. Prescott, A. Ehrenfeucht, and G. Rozenberg. Molecular operations for DNA processing in hypotrichous ciliates. *European Journal of Protistology*, 37:241–260, 2001.

[23] D.J.A. Welsh. *Matroid theory.* Academic Press, 1976.