DOI 10.1007/s10958-020-04928-5
Journal of Mathematical Sciences, Vol. 249, No. 2, August, 2020

DOUBLE OCCURRENCE WORDS: THEIR GRAPHS AND MATRICES

A. E. Guterman,* E. M. Kreines,* and N. V. Ostroukhova† UDC 512.548, 519.177

Double occurrence words play an important part in genetics in describing epigenetic genome rearrangements. A useful geometric representation for double occurrence words is provided by the so-called assembly graphs. The paper investigates properties of the incidence matrices that correspond to the assembly graphs. An explicit matrix characterization of the simple assembly graphs of a given structure and a series of constructions, using these graphs and important for genetic investigations, are provided. Bibliography: 10 *titles.*

1. Introduction

Double occurrence words and the corresponding simple assembly graphs are important notions of branches of algebra and geometry having nontrivial connections. There are two directions of investigations that have originated from applications of these notions. The first one is investigation of numerical invariants of double occurrence words, such as assembly number, minimal realization number, genus, etc. The second one is related to different constructions that allow one to build new families of simple assembly graphs based on already known ones.

In this paper, we consider finite graphs $\Gamma = (V, E)$, where V is the vertex set and $E \subseteq V \times V$ is the edge set. Both loops and multiple edges are allowed.

Definition 1.1. The *degree* or *valency* of a vertex $v \in V$ is the number of edges incident to it. If an edge is twice incident to a vertex (this edge is called a loop), then it is counted twice.

Definition 1.2 ([1, p. 3022])**.** The cyclic order for a k-tuple $(x_1, x_2, x_3, \dots, x_{k-1}, x_k)$ is the set

$$(x_1, x_2, x_3, \dots, x_{k-1}, x_k)^{\mathrm{cyc}} = \{(x_1, x_2, x_3, \dots, x_{k-1}, x_k),\ (x_2, x_3, \dots, x_{k-1}, x_k, x_1),$$
$$(x_3, \dots, x_{k-1}, x_k, x_1, x_2),\ \dots,\ (x_k, x_1, x_2, x_3, \dots, x_{k-1}),\ (x_k, x_{k-1}, x_{k-2}, \dots, x_2, x_1),$$
$$(x_{k-1}, x_{k-2}, \dots, x_2, x_1, x_k),\ (x_{k-2}, \dots, x_2, x_1, x_k, x_{k-1}),\ \dots,\ (x_1, x_k, x_{k-1}, x_{k-2}, \dots, x_2)\},$$

i.e., the set of all cyclic shifts of the tuple and all cyclic shifts of this tuple written in reverse order.

One element of the set $(x_1, x_2, x_3, \dots, x_{k-1}, x_k)^{\mathrm{cyc}}$ is sufficient for specifying the cyclic order because all other elements are obtained as its cyclic shifts and reversed cyclic shifts.

Definition 1.3. A vertex v is said to be *rigid* (or sometimes *regular*) if a cyclic order of the edges incident to this vertex is fixed.

Remark 1.4. 1. If, for example, a graph is embedded into an oriented surface, then all its vertices are rigid.

2. If a vertex is rigid, then for each of its edges the *neighbors* are well defined.

A regular vertex of valency n is also said to be n-regular.

Example 1.5. In Fig. 1, a rigid vertex of degree 4 with the cyclic order of edges (e_1, e_2, e_3, e_4) is shown. It can readily be seen that e_2 and e_4 are the neighbors of e_1 (or e_3) in v.

*Lomonosov Moscow State University, Moscow, Russia and Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia, e-mail: guterman@list.ru, elena.kreines@gmail.com.

†Lomonosov Moscow State University, Moscow, Russia, e-mail: natosova@gmail.com.

Translated from *Zapiski Nauchnykh Seminarov POMI*, Vol. 482, 2019, pp. 45–72. Original article submitted October 7, 2019.

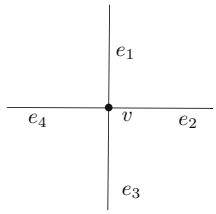

Fig. 1. A rigid vertex of degree 4.

In [5], a special class of rigid graphs, called assembly graphs, is investigated. These graphs appear in genetics and are used in describing epigenetic genome rearrangements.

Definition 1.6. An *assembly graph* is a finite connected graph all of whose vertices are rigid and have valency 1 or 4.

Definition 1.7. Vertices of degree 1 are called *endpoints*.

Definition 1.8. The number of 4-regular vertices in an assembly graph Γ is called the *size* of Γ and is denoted by $|\Gamma|$.

Definition 1.9. An assembly graph is said to be *trivial* if $|\Gamma| = 0$.

Definition 1.10. *Assembly graphs* $\Gamma_1 = (V_1, E_1)$ and $\Gamma_2 = (V_2, E_2)$ are said to be *isomorphic* if $|\Gamma_1| = |\Gamma_2|$ and there exists an isomorphism $\phi : V_1 \to V_2$ such that
(i) for arbitrary $u, v \in V_1$, $(u, v) \in E_1$ if and only if $(\phi(u), \phi(v)) \in E_2$;
(ii) for an arbitrary $u \in V_1$, the cyclic order of edges at u coincides with that of their ϕ-images at $\phi(u)$.

Note that two graphs can be isomorphic as abstract graphs but nonisomorphic as assembly graphs.

Example 1.11. Consider graphs Γ_A and Γ_B presented in Fig. 2. The order of edges and the neighborhood relations can be seen from the figure. For example, the order of edges of the graph Γ_A at vertex 2 is $(a, b, c, d)^{\text{cyc}}$, whereas the order of edges of the graph Γ_B at vertex 2 is $(a, c, b, d)^{\text{cyc}}$. Obviously, these graphs are isomorphic as general graphs. However, this isomorphism does not preserve the cyclic order of edges at vertex 2. For the graph Γ_A, the edges d and b are neighbors of a at vertex 2, but, for the graph Γ_B, the edge a is a neighbor of c and d at vertex 2.

(A) Graph Γ_A (B) Graph Γ_B

Fig. 2. Assembly graphs that are isomorphic as general graphs but not as assembly graphs.

Definition 1.12. A *path* from a vertex u to a vertex v in Γ is a sequence of vertices $u, w_1, \dots, w_l$, $v \in V(\Gamma)$ and a sequence of edges $(u, w_1), (w_1, w_2), \dots, (w_l, v) \in E(\Gamma)$, where the vertices and edges are not necessarily distinct. A path without coinciding vertices is said to be *simple*. A path containing only one vertex is called a *singleton*.

Definition 1.13. A *transverse path*, or a *transversal* is a path in Γ such that all its edges are pairwise distinct and consecutive edges are not neighbors at their common vertex in the sense of the above introduced edge order at vertices.

Definition 1.14. A transversal is said to be *Eulerian* if it passes through all edges of Γ.

Definition 1.15. An assembly graph having an Eulerian transversal is called a *simple assembly graph*.

Definition 1.16. A path in which all vertices are pairwise distinct and all consecutive edges are neighbors at their common vertex is said to be *polygonal*.

Lemma 1.17. *Let Γ be a simple assembly graph with two endpoints, $|\Gamma| = n$. Then Γ contains $2n + 1$ edges. In the case where Γ has no endpoints, it contains $2n$ edges.*

Proof. Since a simple assembly graph has an Eulerian transversal, the number of vertices of degree 1 is either two or zero, whereas all other vertices have degree 4. Therefore, we can count all edges. We take the sum of the valencies of all vertices and divide it by 2 because every edge is incident to two vertices. In the case of two endpoints, the number of edges is $\frac{4 \cdot n + 2}{2} = 2n + 1$, and, in the case of no endpoints, the number of edges equals $\frac{4 \cdot n}{2} = 2n$. □

Given an Eulerian transverse path in a simple assembly graph Γ, $|\Gamma| = n$, one can fix an orientation of this path and in this way obtain an oriented (or a directed) simple assembly graph. Obviously, in a simple assembly graph, there are either two or no endpoints. If there are two endpoints, then an oriented Eulerian transversal starts at one of them (we denote this vertex by 0) and terminates at the other one (we denote it by $n + 1$).

Definition 1.18. Two transversals are said to be *equivalent* if they coincide or one is the reverse of the other.

Below, unless otherwise specified, all graphs are simple assembly graphs with exactly two endpoints.

The assembly graphs are naturally related to a special class of words.

Definition 1.19. An *assembly* word or a *double occurrence* word is a word in a certain alphabet $S = \{a_1, a_2, \dots\}$ such that every symbol a_i either occurs in the word exactly twice or does not occur at all.

Definition 1.20. The word $w^R = a_{i_k} \dots a_{i_1}$ is said to be *reverse* to the word $w = a_{i_1} \dots a_{i_k}$.

Definition 1.21. Two double occurrence words are said to be *equivalent* if, upon renaming some letters, they coincide or are mutually reverse.

Example 1.22. The word $w = 123321$ is equivalent to its reverse, whereas the word $w' = 213132$ coincides with w^R upon interchanging 1 and 3.

We denote the empty double occurrence word by ϵ.

The following theorem interrelates the assembly graphs and the double occurrence words.

Theorem 1.23 ([1, Lemma 3.8])**.** *The equivalence classes of double occurrence words are in one-to-one correspondence with the isomorphism classes of simple assembly graphs.*

The following example illustrates this theorem.

Example 1.24. The assembly words of the graphs Γ_A and Γ_B from Example 1.11, which are isomorphic as abstract graphs but nonisomorphic as assembly graphs, are 1123345542 and 1123443552, respectively. It is straightforward to check that these words are not equivalent.

In order to obtain a word from the graph, one can use the following algorithm.

Algorithm 1.25.

(1) Mark all 4-valent vertices of Γ by the integers $1, \dots, |\Gamma|$.
(2) Choose a vertex of degree 1 as the beginning of the transversal; this vertex is said to be initial.
(3) Follow the transverse path, starting from the initial vertex, and write down the number of every 4-valent vertex each time it is visited by the transversal.

Since every marked vertex has degree 4 and, by definition, a transversal path visits all edges, every 4-valent vertex will be visited twice, whence a double occurrence word will be obtained.

Assume that $\omega = w_1 w_2 \dots w_{2n-1} w_{2n}$ is a double occurrence word. Here, w_i denotes the ith letter of the word, whence w_i are not necessarily distinct. In order to display the corresponding graph, one can use the following algorithm.

Algorithm 1.26.

(1) Draw the initial vertex and start the edge corresponding to it.
(2) Add a vertex to the end of the edge drawn, label it with w_1, and draw an edge outgoing from w_1. Note that so far w_1 has valency 2.
(3) Each of the letters w_i, $i \in \{2, 3, \dots, 2n-1\}$, is processed in the following way:
   (a) After the previous step, we have obtained a graph, a vertex w_{i-1}, and an edge going out of this vertex; the second end of the latter edge is so far not determined.
   (b) In the case where the letter w_i has not previously occurred in the word ω, we draw a new vertex on the edge outgoing from the vertex w_{i-1} and label it with the letter w_i. Then draw an edge outgoing from w_i and go to the beginning of step 3. Note that now the vertex w_i has valency 2, which implies that in the word ω there are unprocessed letters.
   (c) In the case where the letter w_i has already occurred in the word ω, by construction, there is a vertex v_k with valency 2 labeled with the letter w_i. Denote the edges incident to v_k by e_1^k, e_2^k. In this case, we connect the vertex w_{i-1} with v_k, so that the new edge e_3^k meets the vertex v_k between the edges e_1^k and e_2^k. Then we draw an edge e_4^k going out of the vertex v_k between the edges e_1^k and e_2^k. We choose the side where so far there are no edges between e_1^k and e_2^k. Note that in accordance with Definition 1.2, it does not matter which one of the edge tuples $(e_1^k, e_3^k, e_2^k, e_4^k)$ or $(e_1^k, e_4^k, e_2^k, e_3^k)$ (clockwise, starting from e_1^k) is obtained because both tuples define the same cyclic order of edges at v_k.
   (d) If in the word ω there remain unprocessed letters, then go to the beginning of step 3 else proceed to the next step.
(4) When the last letter of the word ω has been processed (obviously, it is the second occurrence of this letter), we complete the drawn "cross" of edges with the terminal vertex.

In Fig. 3, the process of constructing the assembly graph corresponding to the word 1221 is shown.

Given a double occurrence word w, we denote the assembly graph corresponding to this word by Γ_w. The empty word ϵ corresponds to two vertices 0 and 1 connected by an edge.

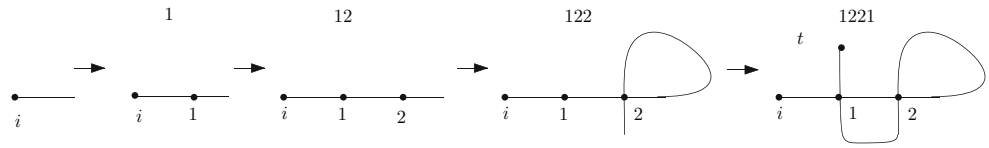

Fig. 3. Construction of a simple assembly graph from the word 1221.

Definition 1.27. A *composition* $\Gamma_1 \circ \Gamma_2$ of two oriented simple assembly graphs Γ_1 and Γ_2 is the graph obtained by identifying the terminal vertex of Γ_1 with the initial vertex of Γ_2 and then forgetting this vertex.

Remark 1.28. It is straightforward to see that a composition of simple assembly graphs itself is a simple assembly graph.

A composition of two words is given by their concatenation. From the definition it immediately follows that given assembly words w_1 and w_2, one has $\Gamma_{w_1} \circ \Gamma_{w_2} = \Gamma_{w_1 w_2}$.

In general, the graphs $\Gamma_1 \circ \Gamma_2$ and $\Gamma_2 \circ \Gamma_1$ are not isomorphic. Consider, for example, the graphs Γ_{aa} and Γ_{bbcddc}. We have $\Gamma_1 \circ \Gamma_2 = \Gamma_{aabbcddc}$, whereas $\Gamma_2 \circ \Gamma_1 = \Gamma_{bbcddcaa}$.

Definition 1.29. The composition $\underbrace{\Gamma \circ \Gamma \circ \cdots \circ \Gamma}_{k}$ is called the kth power of the graph Γ and is denoted by Γ^k.

Now we introduce the notions of Hamiltonian set of polygonal paths and of assembly number, which are the main objects of our investigations. These characteristics of assembly graphs were considered in detail in [1, 2, 5].

Definition 1.30. Two paths are said to be *disjoint* if they have no vertices in common.

We consider disjoint polygonal paths that cover all vertices of an assembly graph.

Definition 1.31. A set $\{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ of pairwise disjoint polygonal paths in Γ is said to be Hamiltonian if their union covers all 4-regular vertices of Γ.

For example, the set $V(\Gamma)$ of all vertices is a Hamiltonian set of singletons.

Definition 1.32. A polygonal path γ is said to be Hamiltonian if the set $\{\gamma\}$ is Hamiltonian.

Let Γ be a nontrivial assembly graph.

Definition 1.33. The *assembly number* of Γ (denoted by $\mathrm{An}(\Gamma)$) is the smallest size of a Hamiltonian set of polygonal paths, i.e.,

$$\min\{k : \text{there exists a Hamiltonian set of polygonal paths } \{\gamma_1, \gamma_2, \ldots, \gamma_k\} \text{ in } \Gamma\}.$$

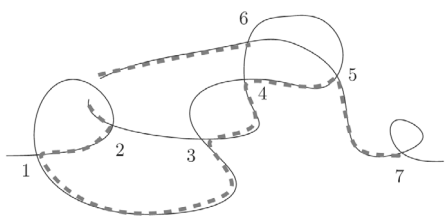

Fig. 4. The realizable graph with assembly word 12134564326577 and a Hamiltonian polygonal path (shown by the dashed line).

Definition 1.34. A graph with $\mathrm{An}(\Gamma) = 1$ is said to be realizable; otherwise it is said to be unrealizable.

Definition 1.35. For a positive integer n, the minimal realization number is defined as the smallest size of an assembly graph with assembly number n: $\mathrm{R}_{\min}(n) = \min\{|\Gamma| : \mathrm{An}(\Gamma) = n\}$.

It is well known that assembly graphs are used in describing epigenetic genomic rearrangements, see, for example, [6]. Following [1], we briefly describe this process here for the sake of completeness; for more detail, see [1, 6] and the references therein.

There are two types of nuclei, micronuclear and macronuclear, and both can be represented in several copies. Only micronuclear genes are exchanged during mating. After conjugation, the old macronuclei are destroyed, and new macronuclei are constructed from one of the newly formed micronuclei. These DNA processing events involve effective destruction of all the so-called "junk" DNA, intervening DNA segments (internal eliminated sequences, IESs) that interrupt coding of the genes. Note that the newly constructed DNA contains 95–98% of intervening segments. Since IESs interrupt coding regions in the micronucleus, every macronuclear gene may appear as several nonconsecutive segments (macronuclear destined sequences, MDSs) in the micronucleus. Moreover, for thousands of genes, even the order of these MDS segments in the micronuclei can be permuted, or sequences can be reversed with respect to the micronuclear sequence.

There are several theoretical models attempting to describe these DNA recombination processes [6–8, 10]. It has been conjectured that an additional molecule takes part in the recombination process [2, 10], and experimental support for this model was obtained in [9].

Based on these observations, a theoretical model with spatial graphs depicting the molecule(s) at the time of recombination was introduced in [2]. This model describes a micronuclear gene as a graph with 1- or 4-valent regular vertices. Every 4-valent vertex represents the location of the homologous recombination. A single micronuclear gene is modeled as an assembly graph with an Eulerian path, in which consecutive edges are not "neighbors" with respect to the common incident vertex. Observe that the sequence of vertices listed in the order they are visited by an Eulerian path is an assembly word.

The aim of this paper is to describe simple assembly graphs in terms of their incidence matrices. In particular, we characterize matrices corresponding to several standard series of graphs and translate into matrix language certain important procedures on graphs, which are actual in genetic applications. In particular, we characterize procedures of loop addition and graph concatenation.

The paper is organized as follows. In Sec. 2, we collect basic properties of the incidence matrices of assembly graphs. Section 3 contains some examples of the incidence matrices of some special series of assembly graphs. In Secs. 4 and 5, we describe, in matrix terms, several standard procedures, actively used in modifying assembly graphs, such as (interior) loop saturation and composition, which are important for genetics.

2. Incidence matrices

In this section, we recall the definition of the incidence matrix of an arbitrary graph and describe some properties of the incidence matrices of simple assembly graphs.

Definition 2.1. Let Γ be a general (not necessarily simple assembly) graph of order n with vertices $v_1, \ldots, v_n$ and edges $e_1, e_2, \ldots, e_m$. The incidence matrix of Γ is the $n \times m$ integral

matrix $I(\Gamma) = (a_{ij})$ defined by

$$a_{ij} = \begin{cases} 1 & \text{if } v_i \text{ is incident to } e_j \text{ and } e_j \text{ is not a loop;} \\ 2 & \text{if } v_i \text{ is incident to } e_j \text{ and } e_j \text{ is a loop;} \\ 0 & \text{otherwise.} \end{cases}$$

Remark 2.2. Let $\Gamma = (V, E)$, $|\Gamma| = n$, be a simple assembly graph. We can enumerate its vertices and edges in ascending order corresponding to the transversal path from the initial vertex to the terminal one starting from 0. By Lemma 1.17, we have $V = \{0, 1, \dots, n, n+1\}$ and $E = \{e_0, e_1, \dots e_{2n-1}, e_{2n}\}$.

Definition 2.3. The incidence matrix of a simple assembly graph Γ is the incidence matrix of Γ, where the order of vertices and edges is fixed as in Remark 2.2.

Example 2.4. In Fig. 5, one can see the graph with assembly word $w = 1122$ and edges labeled in the order we meet them following the transversal.

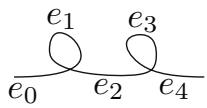

Fig. 5. The graph with assembly word 1122.

Below, we present the incidence matrix of the graph from Fig. 5.

$$I(\Gamma_{1122}) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} e_0 & e_1 & e_2 & e_3 & e_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Proposition 2.5. *The incidence matrices of simple assembly graphs possess the following properties*:

(1) *The incidence matrix of a simple assembly graph with* $|\Gamma| = n$ *has* $n+2$ *rows and* $2n+1$ *columns.*

(2) *For all rows except for the first and last ones, the row sums of entries equal* 4. *The row sums of the first and last rows equal* 1.

(3) *Two nonconsecutive columns that have a common row with nonzero entries correspond to two edges that are neighbors at their common vertex.*

(4) *Two neighboring columns corresponding to edges* e_k, e_{k+1} *always share at least one common row with nonzero entries.*

Proof. (1) This assertion immediately follows from Lemma 1.17.

(2) Since all vertices of a simple assembly graph are 4-valent, except for the initial and terminal ones, the result follows.

(3) If a row has two nonzero entries, then the corresponding edges are incident to the same vertex. We label edges following the transversal, which implies that consecutive columns correspond to edges that are not neighbors at their common vertex. For a fixed vertex v and an edge e incident to it, there are exactly two neighbors of e at v and only one edge incident to v that is not a neighbor of e at v. Hence two nonconsecutive columns sharing a common row with nonzero entries correspond to neighboring edges.

(4) Indeed, since we write edges in the order we meet them following the transversal, it follows that two consecutive edges always share a common vertex. □

For example, both columns corresponding to edges e_1 and e_2 in the matrix $I(\Gamma_{1122})$ have nonzero entries only in the row corresponding to vertex 1.

3. Structured graphs and their matrices

In this section, we consider series of graphs, which are well known in many examples due to their extremal behavior, see [1, 2, 5]. We start with very simple examples of assembly graphs.

3.1. Examples of simple assembly graphs. We start with a graph free of 4-valent vertices. Then we consecutively add loops and obtain the graph sequence $\Gamma_\epsilon, \Gamma_{11}, \Gamma_{1122}, \Gamma_{112233}$, presented in Fig. 6.

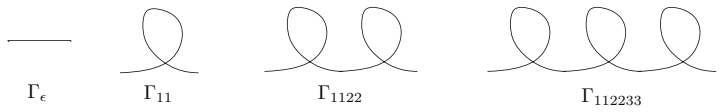

Fig. 6. Realizable assembly graphs.

The incidence matrices of these graphs are as follows:

$$I(\Gamma_\epsilon) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} e_0 \\ \left(\begin{array}{c} 1 \\ 1 \end{array}\right) \end{array}, \quad I(\Gamma_{11}) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} e_0 \quad e_1 \quad e_2 \\ \left(\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 2 & 1 \\ 0 & 0 & 1 \end{array}\right) \end{array},$$

$$I(\Gamma_{1122}) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} e_0 \quad e_1 \quad e_2 \quad e_3 \quad e_4 \\ \left(\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array},$$

$$I(\Gamma_{112233}) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} e_0 \quad e_1 \quad e_2 \quad e_3 \quad e_4 \quad e_5 \quad e_6 \\ \left(\begin{array}{ccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}.$$

As is readily seen, a row corresponding to $v_k, k \in \{1, \ldots, n\}$, has a block 121 that starts from the column corresponding to $e_{2(k-1)}$.

3.2. Graphs with given assembly numbers. Show how one can construct a graph with an arbitrary given assembly number. Consider the assembly graph with assembly word $u = 122133$ (see Fig. 7).

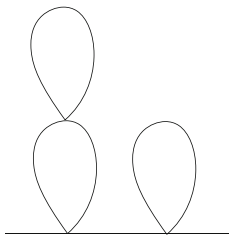

Fig. 7. The assembly graph Γ_u with assembly word $u = 122133$.

Its incidence matrix is

$$I(\Gamma_u) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccccc} e_0 & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Now consider the composition $\Gamma_n := \Gamma_{aa} \circ \Gamma_u^n$. As was shown in [1], this composition has assembly number $\mathrm{An}(\Gamma_n) = n + 1$. Therefore, the graph Γ_u allows us to construct an example of an assembly graph with given assembly number k. Note that $|\Gamma_k| = 3k + 1$ by construction.

Figures 8 and 9 present the graphs with assembly numbers 2 and 3 constructed as described above. In Sec. 5, the process of finding the incidence matrix of a composition of two graphs is described in more detail.

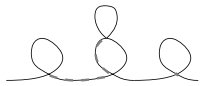

Fig. 8. The assembly graph $\Gamma_{aa} \circ \Gamma_u$ with assembly number 2.

The incidence matrix of $\Gamma_{aa} \circ \Gamma_u$ is as follows:

$$I(\Gamma_{aa} \circ \Gamma_u) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} e_0 & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Fig. 9. The assembly graph $\Gamma_{aa} \circ \Gamma_u^2$ with assembly number 3.

The incidence matrix of the graph $\Gamma_{aa} \circ \Gamma_u^2$ is

$$I(\Gamma_{aa} \circ \Gamma_u^2) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{ccccccccccccccc} e_0 & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 & e_9 & e_{10} & e_{11} & e_{12} & e_{13} & e_{14} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

3.3. Tangled cord

Definition 3.1. Let n be a positive integer. A tangled cord of size n, denoted by TC_n, is the simple assembly graph corresponding to the word defined inductively in the following way: The starting word is $w_{TC_1} = 11$, and w_{TC_n} is obtained from $w_{TC_{n-1}}$ by replacing the last letter $n-1$ with the subword $n(n-1)n$.

For example, the first four words are 11, 1212, 121323, 12132434; the nth word is $w_{TC_n} = 121324354\cdots(n-1)(n-2)n(n-1)n$.

Figures 10, 11, and 12 below provide examples of such graphs for $n = 2, 3, 4$, along with the corresponding words and matrices.

Example 3.2. Let $n = 2$. Then $w_{TC_2} = 1212$,

$$I(TC_2) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} e_0 & e_1 & e_2 & e_3 & e_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array},$$

and the corresponding graph is presented in Fig. 10.

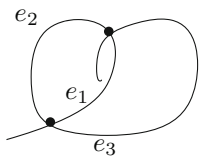

Fig. 10. The graph TC_2 with labeled edges.

Example 3.3. Let $n = 3$. Then $w_{TC_3} = 121323$,

$$I(TC_3) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccccc} e_0 & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array},$$

and the corresponding graph is presented in Fig. 11.

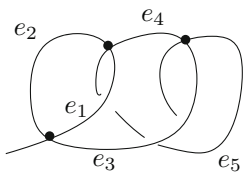

Fig. 11. The graph TC_3 with labeled edges.

Example 3.4. Let $n = 4$. Then $w_{TC_4} = 12132434$,

$$I(TC_4) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} e_0 & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array},$$

and the corresponding graph is presented in Fig. 12.

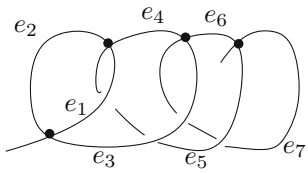

Fig. 12. The graph TC_4 with labeled edges.

Figure 13 shows the general form of a tangled cord and the process of adding vertices.

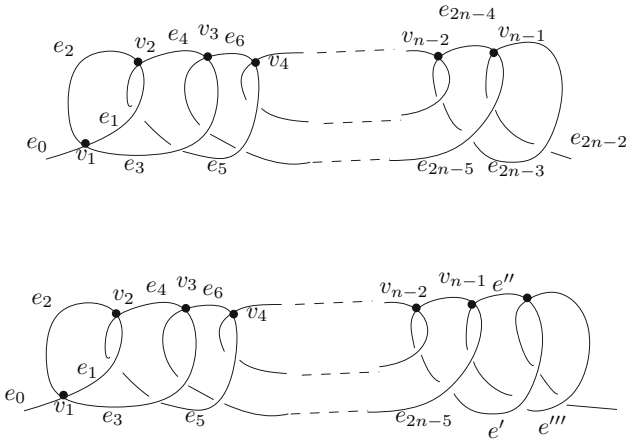

Fig. 13. Adding a vertex to TC_{n-1}.

Proposition 3.5. *For an arbitrary $n \geq 3$, the rows of the matrix $I(TC_n)$ have the following form:*

$$R_0(TC_n) = (1, \underbrace{0, \ldots, 0}_{2n \text{ times}}), \quad R_1(TC_n) = (1, 1, 1, 1, \underbrace{0, \ldots, 0}_{2n-3 \text{ times}}),$$
$$R_k(TC_n) = (\underbrace{0, \ldots, 0}_{2k-3 \text{ times}}, 1, 1, 0, 1, 1, \underbrace{0, \ldots, 0}_{2n-2k-1 \text{ times}}), \quad k = 2, \ldots, n-1;$$
$$R_n(TC_n) = (\underbrace{0, \ldots, 0}_{2n-3 \text{ times}}, 1, 1, 1, 1), \quad R_{n+1}(TC_n) = (\underbrace{0, \ldots, 0}_{2n \text{ times}}, 1).$$

Proof. It is straightforward to see that the row corresponding to vertex 0 (resp., $n+1$) contains only one entry 1 in the first (resp., last) column, whereas the other entries of these rows are zero. The row corresponding to vertex 1 (resp., n) contains ones in the four consecutive columns starting from e_0 (resp., starting from $e_{(2n-1)-2}$). Any other row, corresponding to a vertex $k \in \{2, \dots, n-1\}$, contains the block [1 1 0 1 1], which starts from the column corresponding to e_{2k-3}. The proof is completed by counting the entries. □

3.4. Return words

Definition 3.6. A *return word* is a double occurrence word of the form

$$a_1 a_2 \dots a_{n-1} a_n a_n a_{n-1} \dots a_2 a_1.$$

In what follows, a return word with n different letters will be denoted by $\mathrm{ret}(n)$.

Return words are used to represent the parts of the micronuclear genome corresponding to frequently occurring sequences and play an important part in studying the nesting index, see [3, 4] for more detail.

Example 3.7. Consider, for example, the double occurrence word 12344321. The corresponding graph is shown in Fig. 14.

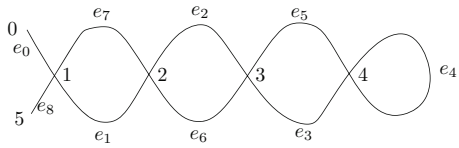

Fig. 14. The assembly graph corresponding to ret(4).

The incidence matrix of this graph is

$$I(\Gamma_{\mathrm{ret}(4)}) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} e_0 & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Proposition 3.8. *For any $n \geqslant 2$, the rows of the matrix $I(\Gamma_{\mathrm{ret(n)}})$ have the following form:*

$$R_0(\mathrm{ret}(n)) = (1, \underbrace{0, \dots, 0}_{2n \text{ times}}),$$

$$R_k(\mathrm{ret}(n)) = (\underbrace{0, \dots, 0}_{k-1 \text{ times}}, 1, 1 \underbrace{0, \dots, 0}_{2n-2k-1 \text{ times}}, 1, 1, \underbrace{0, \dots, 0}_{k-1 \text{ times}}), \quad k = 1, \dots, n-1;$$

$$R_n(\mathrm{ret}(n)) = (\underbrace{0, \dots, 0}_{n-1 \text{ times}}, 1, 2, 1, \underbrace{0, \dots, 0}_{n-1 \text{ times}}),$$

$$R_{n+1}(\mathrm{ret}(n)) = (\underbrace{0, \dots, 0}_{2n \text{ times}}, 1).$$

In other words, the incidence matrix (excluding the first and last rows) of the graph with assembly word $\mathrm{ret}(n)$ *is symmetric with respect to the column labeled by* e_n.

Proof. The first and last rows correspond to the initial and terminal vertices, having only one edge. Therefore, in each of these rows, there is only one nonzero entry. The only vertex with a loop is vertex n, and the loop is in the middle of the Eulerian transversal, whence it is e_n. Obviously, e_{n-1} and e_{n+1} are also incident to n. The return word is a palindrome, which implies that nothing changes as we follow the Eulerian transversal backwards. Hence e_k and e_{2n-k} connect the same vertices. Obviously, every edge e_k, $k = 0, 1, \ldots, n-1$, connects vertices with numbers $k, k+1$. □

3.5. Repeat words

Definition 3.9. A *repeat word* is a double occurrence word of the form

$$a_1 a_2 \ldots a_{n-1} a_n a_1 a_2 \ldots a_{n-1} a_n.$$

Note that any repeat word with n letters is equivalent to $123\ldots n123\ldots n$. Thus, in the sequel, we denote the repeat word with n different letters by $\mathrm{rep}(n)$.

Along with return words, repeat words help us in studying the complexity of the micronuclear gene, see [3, 4] for more detail.

Example 3.10. Consider, for example, the word $\mathrm{rep}(4) = 12341234$. The corresponding graph is shown in Fig. 15.

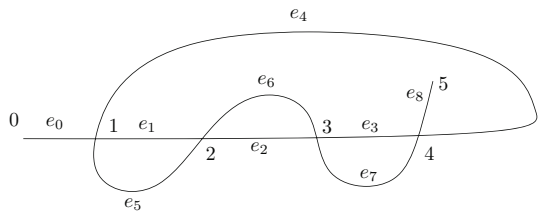

Fig. 15. The assembly graph corresponding to rep(4).

The incidence matrix of this graph is as follows:

$$I(\Gamma_{\mathrm{rep}(4)}) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} e_0 & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Proposition 3.11. *For any $n \geq 2$, the rows of the matrix $I(\Gamma_{\mathrm{rep}}(n))$ are as follows:*

$$R_0(\mathrm{rep}(n)) = (1, \underbrace{0, \ldots, 0}_{2n \text{ times}}),$$

$$R_k(\mathrm{rep}(n)) = (\underbrace{0, \ldots, 0}_{k-1 \text{ times}}, 1, 1\ \underbrace{0, \ldots, 0}_{n-2 \text{ times}}, 1, 1,\ \underbrace{0, \ldots, 0}_{n-k \text{ times}}), \quad k = 1, \ldots, n;$$

$$R_{n+1}(\mathrm{rep}(n)) = (\underbrace{0, \ldots, 0}_{2n \text{ times}}, 1).$$

Proof. Consider the chain of edges $e_0, \ldots, e_{n-1}$, corresponding to the first n columns of our matrix. Every edge $e_i, i \in \{0, 1, \ldots, n-1\}$, connects vertices i and $i+1$. This implies that the entries $R_i(\mathrm{rep}(n))[i]$ (the ith entry of the ith row) and $R_{i+1}(\mathrm{rep}(n))[i]$ (the ith entry of the

$(i+1)$st row) equal 1 for $i \in \{0, 1, \ldots, n-1\}$. The edge e_n connects vertices n and 1, whence $R_1(\mathrm{rep}(n))[n] = 1$ and $R_n(\mathrm{rep}(n))[n] = 1$. Consider the edges $e_{n+1}, \ldots, e_{2n}$, corresponding to the last n columns. The edge $e_i, i \in \{n+1, n+2, \ldots, 2n\}$, connects vertices $i-n$ and $i-n+1$. Therefore, the entries $R_i(\mathrm{rep}(n))[n+i]$ and $R_{i+1}(\mathrm{rep}(n))[n+i]$ equal 1 for $i \in \{n+1, n+2, \ldots, 2n\}$. Thus, the incidence matrix of our graph has the form indicated. □

4. Interior Loop Saturation

Interior loop saturation is an important transformation of assembly graphs originating from genetics, see [5] for details. In this section, we describe the corresponding matrix transformations. We start with the necessary definitions.

Definition 4.1. A *loop* is the assembly graph with assembly word 11, see Fig. 16.

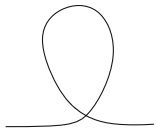

Fig. 16. The loop $\Gamma_{(1)}$.

The next lemma is obvious.

Lemma 4.2 ([5, p. 15, Definition 5.1])**.** *Let Γ be an assembly graph and let Γ' be obtained from Γ by replacing one of the edges with the loop $\Gamma_{(1)}$. Then Γ' is an assembly graph.*

Proof. As is readily seen, all vertices of Γ' have degree 1 or 4, and the number of vertices with degree 1 remains unchanged. Show that Γ' has an Eulerian transversal. Consider the Eulerian transversal of Γ. Since the edge replaced by a loop is a part of this transversal, it is possible to include the loop into the existing Eulerian transversal and obtain the desired path in Γ'. □

Figure 17 shows how a loop is incorporated into a transverse path in the two standard cases where a loop is adjoined to an edge that is a loop and to an edge that is not a loop.

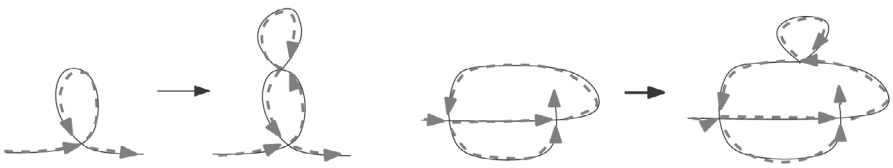

Fig. 17. Incorporating a loop into an existing transversal path.

Applying this lemma a number of times, we see that upon replacing some edges of an assembly graph by loops, an assembly graph is obtained. This leads us to the following definition.

Definition 4.3 ([5, Definition 5.1])**.** Let Γ be an assembly graph and let $\tilde{\Gamma}$ be obtained from Γ by adding the loop $\Gamma_{(1)}$ to every edge. Then $\tilde{\Gamma}$ is called a *loop-saturated graph* or a graph obtained by loop-saturation. An assembly graph $\tilde{\Gamma}^\circ$ is called an *interior loop-saturated graph* if it is obtained from Γ by adding the loop $\Gamma_{(1)}$ to every edge, except for two edges incident to the endpoints.

The incidence matrix of the graph obtained by loop-saturation (or interior loop-saturation) can readily be constructed from the incidence matrix of a given graph.

Proposition 4.4. *Let Γ be an assembly graph and let $\Gamma^{(i)}$ be obtained from Γ by changing the ith edge for the loop $\Gamma_{(1)}$. Then the following assertions hold:*

1. *$I(\Gamma^{(i)})$ is obtained from $I(\Gamma)$ by replacing the ith column c_i with three columns C_1, C_2, C_3 and adding a row R as described below. Let the nonzero entries of c_i be located in the kth and lth rows, $k \le l$. If $k = l$, then R is inserted below row k. If $k \ne l$, then R is inserted above row l. The entries of the inserted columns are as follows: C_2 contains 2 in the row R; C_1 contains 1 in the row R and 1 in the kth row; C_3 contains 1 in the row R and 1 in the former lth row. All the other entries of the inserted rows and columns are zero.*

2. *$I(\Gamma)$ is obtained from $I(\Gamma^{(i)})$ by removing the row and column corresponding to $\Gamma_{(1)}$ and substituting two columns with 1 in the removed row by their sum.*

Proof. 1. Adjoining a loop to a graph results in that two edges and a vertex appear. By definition, the loop corresponds to a column with 2 in the new row. The edge between k and l is split by the new vertex into two new edges. Therefore, the corresponding column must be changed for the following two columns: The column C_1 contains unit entries in row k and the new row. The column C_2 contains unit entries in the former lth row and the new row. Also, the column C_2, corresponding to the loop, with entry 2 in the new row, must be inserted between them. All the other entries of C_1, C_2, C_3 are zero.

2. Conversely, if $\Gamma_{(1)}$ is the loop on the ith edge, then the row corresponding to the vertex of $\Gamma_{(1)}$ contains exactly three nonzero entries: 1, 1, 2. It is necessary to delete the column with entry 2 and change two other columns either for the column with ones in the same rows where they used to be (if these rows are different) or for the column with 2 in the row in which the columns being replaced have unit entries. Obviously, this is exactly the sum of the columns removed. □

Corollary 4.5. *Let Γ be an assembly graph. The incidence matrix $I(\tilde{\Gamma})$ of the loop saturated (resp., the incidence matrix $I(\tilde{\Gamma}^\circ)$ of the interior loop saturated) graph can be obtained from $I(\Gamma)$ by successively applying the procedure described in Proposition* 4.4 *to every column (resp., every column except the first and last ones). Similarly, the converse procedure can be applied. Note that if $|\Gamma| = n$, then $I(\Gamma) \in M_{n+2,2n+1}$, $I(\tilde{\Gamma}) \in M_{3n+3,6n+3}$, and $I(\tilde{\Gamma}^\circ) \in M_{3n+1,6n-1}$.*

Example 4.6. Let Γ be the graph shown in Fig. 18. Then Γ is the assembly graph related to

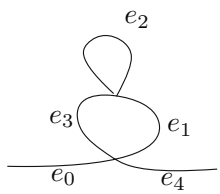

Fig. 18. The graph $\Gamma = \Gamma_{1221}$ with labeled edges.

the word $w = 1221$. The corresponding loop saturation graph $\tilde{\Gamma}$ is shown in Fig. 19.

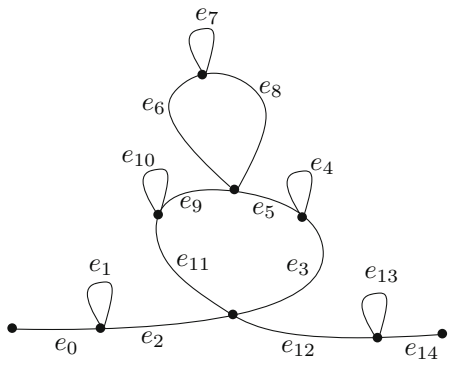

Fig. 19. The loop saturation graph $\tilde{\Gamma}$.

As is straightforward to check, the related incidence matrices are as follows:

$$I(\tilde{\Gamma}) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{ccccccccccccccc} e_0 & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 & e_9 & e_{10} & e_{11} & e_{12} & e_{13} & e_{14} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

and

$$I(\Gamma_{1221}) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} e_0 & e_1 & e_2 & e_3 & e_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Below, we provide an algorithm allowing one to construct the matrix of an original assembly graph from the matrix of its loop saturation graph. In order to simplify the notation, in the next algorithm we assume that the loop saturation graph has $3n+3$ vertices and the interior loop saturation graph has $3n+1$ vertices. In this case, the loop saturation graph has $6n+3$ edges, and the interior loop saturation graph has $6n-1$ edges, respectively.

Algorithm 4.7.

1. Mark all the rows that correspond to loops.
2. Replace any triple of columns with numbers $k, k+1, k+2$ by the column equal to the vector sum of the columns with numbers $k, k+2$. Here, $k = 0, 3, 6, \ldots, 6n$ in the case of loop saturation and $k = 1, 4, 7, \ldots, 6n-5$ in the case of interior loop saturation.
3. Delete the rows marked at step 1.

Proof. The algorithm is based on the fact that every occurrence of 2 in the incidence matrix corresponds to a loop added during loop saturation. Therefore, we delete the rows and columns that contain the entry 2 and sew together columns $k, k+2$ that both belong to the same edge. All the column triples are processed in the same way, starting from the zeroth column in the case of loop saturation and from the first column in the case of interior loop saturation. □

Example 4.8. In order to illustrate the above algorithm, consider the loop saturation graph shown in Fig. 19. First, we mark all the rows that correspond to loops:

$$I(\Gamma) = \begin{array}{c} \\ 0 \\ \mathbf{1} \\ 2 \\ \mathbf{3} \\ 4 \\ \mathbf{5} \\ \mathbf{6} \\ \mathbf{7} \\ 8 \end{array} \begin{array}{c} \begin{array}{ccccccccccccccc} e_0 & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 & e_9 & e_{10} & e_{11} & e_{12} & e_{13} & e_{14} \end{array} \\ \left(\begin{array}{ccccccccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right) \end{array}.$$

Then we perform the second step and replace columns with the corresponding vector sums:

$$\begin{array}{c} \\ 0 \\ \mathbf{1} \\ 2 \\ \mathbf{3} \\ 4 \\ \mathbf{5} \\ \mathbf{6} \\ \mathbf{7} \\ 8 \end{array} \begin{array}{c} \begin{array}{ccccc} e_0+e_2 & e_3+e_5 & e_6+e_8 & e_9+e_{11} & e_{12}+e_{14} \end{array} \\ \left(\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right) \end{array}.$$

Finally, we delete the marked rows and obtain the incidence matrix of the graph with assembly word $w = 2442$,

$$I(\Gamma') = \begin{array}{c} \\ 0 \\ 2 \\ 4 \\ 8 \end{array} \begin{array}{c} \begin{array}{ccccc} e_0+e_2 & e_3+e_5 & e_6+e_8 & e_9+e_{11} & e_{12}+e_{14} \end{array} \\ \left(\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right) \end{array}.$$

Obviously, the incidence matrix $I(\Gamma_{1221})$ of the graph from Fig. 18 and the matrix $I(\Gamma')$ coincide.

5. Incidence matrix of a composition

The incidence matrix of a composition is obtained as a block matrix whose blocks correspond to the incidence matrices of the components, except for the rows corresponding to the terminal (initial) vertices.

Proposition 5.1. *Let Γ_1 and Γ_2 be two simple assembly graphs with incidence matrices $I(\Gamma_1)$ and $I(\Gamma_2)$ and sizes $|\Gamma_1| = n$ and $|\Gamma_2| = m$. Then the incidence matrix of the composition $I(\Gamma_1 \circ \Gamma_2)$, with $n + m + 2$ rows and $2n + 2m + 1$ columns, has the following form (here, the*

rows and columns are enumerated starting from 0):

(1) $I(\Gamma_1 \circ \Gamma_2)_{k,l} = I(\Gamma_1)_{k,l}$ *if* $0 \le k \le n$, $0 \le l \le 2n$;

(2) $I(\Gamma_1 \circ \Gamma_2)_{k,l} = 0$ *if* $n < k \le m+n+1$ *and* $0 \le l < 2n$
or $0 \le k < n+1$ *and* $2n < l \le 2m+2n$;

(3) $I(\Gamma_1 \circ \Gamma_2)_{k,l} = I(\Gamma_2)_{k-n,l-2n}$ *if* $n+1 \le k \le m+n+1$, $2n \le l \le 2n+2m$.

Proof. Indeed, the vertices of the composition graph are divided into two parts: those of Γ_1, except for the terminal vertex, and the vertices of Γ_2, except for the initial vertex. Within each of the parts, the vertices are connected by the same edges as in the original graphs. Vertices from different parts are not connected, except for the edge e_{2n}, corresponding to column $2n$, that connects the two parts. □

We illustrate the above proposition by considering two graphs with assembly words 1212 and 1221 (see Fig. 20).

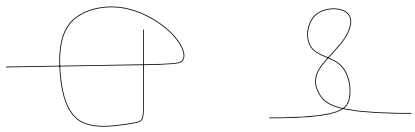

Fig. 20. Graphs before composition.

These graphs have the following incidence matrices:

$$I(TC_2) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} e_0 & e_1 & e_2 & e_3 & e_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}, \quad I(\Gamma_{1221}) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} e_0 & e_1 & e_2 & e_3 & e_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

The incidence matrix of the composition $1212 \circ 1221$ is constructed in the following way. We take $I(TC_2)$ and delete the last row. Then we take $I(\Gamma_{1221})$ and delete the first row. After that we write $I(\Gamma_{1221})$ below, starting with the column that corresponds to the edge incident to the terminal vertex of the first matrix, and fill the remaining positions with zeros:

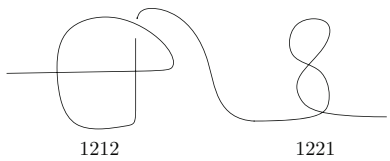

Fig. 21. The graph obtained by composition.

	e_0	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
0	1	0	0	0	0	0	0	0	0
1	1	1	1	1	0	0	0	0	0
2	0	1	1	1	1	0	0	0	0
3	0	0	0	0	1	1	0	1	1
4	0	0	0	0	0	1	2	1	0
5	0	0	0	0	0	0	0	0	1

.

The authors are grateful to Natasha Jonoska for inspiring starting discussions.

The work of the first two authors was supported by the Russian Science Foundation (project No. 17-11-01124).

Translated by N. V. Ostroukhova.

REFERENCES

1. A. Angeleska, N. Jonoska, and M. Saito, "DNA recombinations through assembly graphs," *Discr. Appl. Math.*, **157**, 3020–3037 (2009).
2. A. Angeleska, N. Jonoska, M. Saito, and L. F. Landweber, "RNA-guided DNA assembly," *J. Theor. Biology*, **248** (4), 706–720 (2007).
3. R. Arredondo, "Properties of graphs used to model DNA recombination," *Grad. Theses Diss.* (2014). https://scholarcommons.usf.edu/etd/4979
4. R. Arredondo, "Reductions on double occurrence words," arXiv:1311.3543.
5. J. Burns, E. Dolzhenko, N. Jonoska, T. Muche, and M. Saito, "Four-regular graphs with rigid vertices associated to DNA recombination," *Discr. Appl. Math.*, **161**, 1378–1394 (2013).
6. A. Ehrenfeucht, T. Harju, I. Petre, D. M. Prescott, and G. Rozenberg, *Computing in Living Cell*, Springer (2005).
7. A. Ehrenfeucht, T. Harju, and G. Rozenberg, "Gene assembly through cyclic graph decomposition," *Theor. Comp. Sci.*, **281**, 325–349 (2002).
8. L. Kari and L. F. Landweber, "Computational power of gene rearrangement," in: DNA Based Computers (E. Winfree, D. K. Gifford eds.), AMS (1999), pp. 207–216.
9. M. Nowacki, V. Vijayan, Y. Zhou, K. Schotanus, T. G. Doak, and L. F. Landweber, "RNA-mediated epigenetic programming of a genome-rearrangement pathway," *Nature*, **451**, 153–159 (2008).
10. D. M. Prescott, A. Ehrenfeucht, and G. Rozenberg, "Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates," *J. Theor. Biology*, **222**, 323–330 (2003).