Gene Assembly Through Cyclic Graph Decomposition

A. Ehrenfeucht* T. Harju† G. Rozenberg‡§

Abstract

We present in this paper a graph theoretical model of gene assembly, where (segments of) genes are distributed over a set of circular molecules. This model is motivated by the process of gene assembly in ciliates, but it is more general. In this model a set of circular DNA molecules is represented by a bicoloured and labelled graph γ consisting of cyclic graphs, and the recombination takes place in two stages: first, by folding $\gamma * P$ with respect to a set P of pairs of vertices of the graph (representing pointers in the micronuclear genes of the ciliate), and secondly, by unfolding the so obtained graph to $\gamma \circledast P$ with respect to vertices of higher valency. The final graph $\gamma \circledast P$ is again a set of bicoloured cyclic graphs, where the genes are present as maximal monochromatic paths. Thus, the process of gene assembly corresponds to the dynamic process of changing cyclic graph decompositions. We show that the operation $\circledast$ is well behaved in many respects, and that there is a sequence of pointer sets $P_1, \dots, P_m$ consisting of one or two pairs such that $\gamma \circledast P = (\dots((\gamma \circledast P_1) \circledast P_2) \dots \circledast P_m)$ and each intermediate step $\gamma_i = (\dots((\gamma \circledast P_1) \circledast P_2) \dots \circledast P_i)$ is intracyclic, that is, the segments of a gene that lie in the same connected component of γ_i, will lie in the same connected component of the successor graph γ_{i+1}.

*Department of Computer Science, University of Colorado at Boulder, Boulder, Co 80309-0347, USA – email: andrzej@cs.colorado.edu

†Department of Mathematics, University of Turku, FIN-20014 Turku, Finland – email: harju@utu.fi

‡Leiden Institute for Advanced Computer Science, Leiden University, Leiden, the Netherlands – email: rozenber@liacs.nl

§Department of Computer Science, University of Colorado at Boulder, Boulder, Co 80309-0347, USA

1 Introduction

DNA computing, or more generally molecular computing, is an exciting interdisciplinary research area investigating the use of biomolecules for the purpose of computing. DNA computing *in vivo* investigates computational properties of DNA in its favourite environment: the living cell.

The process of gene assembly in ciliates (very ancient single cell organisms) is a prime example of DNA computing *in vivo*. Ciliates have developed a unique feature of *nuclear dualism* - they have two nuclei that are functionally different: the *micronucleus* and the *macronucleus*. The micronucleus is a germlike nucleus that gets 'activated' only during the process of sexual reproduction, and the macronucleus is a somatic nucleus providing RNA transcripts needed for the vegetative functioning of the cell.

When ciliates are starved, they proceed to sexual reproduction, and during this process micronuclear genes are converted into their macronuclear form. This conversion process is called *gene assembly*. The process of gene assembly is very intricate, because the micronuclear and the macronuclear forms of the same gene may be drastically different. As a matter of fact, the DNA processing in ciliates is among the most sophisticated DNA processing in living organisms. Even more important for us is that the process of gene assembly is fascinating from the computational point of view.

The DNA in the micronucleus is very long (hundreds of thousands base pairs) and it consists mostly of spacer DNA 'interrupted' by genes that occur either individually or in groups. On the other hand, the DNA molecules in the macronucleus are gene-size, on average about 2000 base pairs long. Also, the form of the genes is very different. Genes in the micronucleus are interrupted by multiple noncoding segments called IESs (*Internal Eliminated Segments*). Then, the segments of genes interrupted by IESs are called MDSs (*Macronuclear Destined Segments*) - the structure of a micronuclear gene is given in Fig. 1, where MDSs are given as rectangles (with 'pointers' at the ends) and the interspersing IESs are given by line segments. The same gene in the macronucleus has the form shown in Fig. 2, and so it consists of MDSs from its micronuclear form that are spliced together by 'gluing' them on common 'pointers'.

Figure 1: The structure of a micronuclear gene

More specifically each MDS M_i has the form $M_i = (p_i, \mu_i, p_{i+1})$ except for M_1 which has the form $M_1 = (b, \mu_1, p_2)$, and M_k which has the from $M_k = (p_k, \mu_k, e)$. The double stranded 'boundary' segments p_i are called *pointers*, and the double stranded segments μ_i are called *bodies*. On the other hand, b in M_1 and e in M_k are merely (symbolic) *markers* indicating

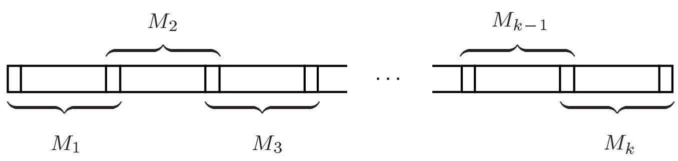


Figure 2: The structure of a macronuclear gene

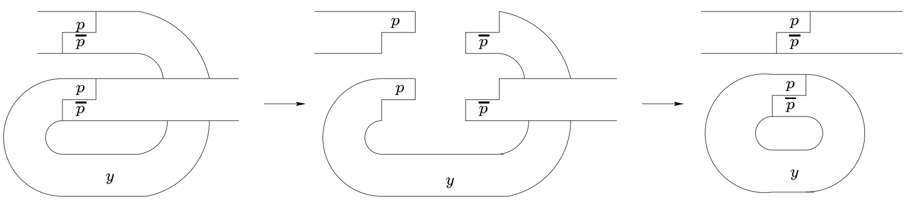


Figure 3: The ld-excision operation.

the beginning and the end of the macronuclear gene (where telomeres will be attached after the gene excision).

Thus, during the process of gene assembly the micronuclear gene of the form given in Fig. 1 will be converted into the macronuclear gene of the form given in Fig. 2. The order of MDSs in the micronuclear gene is a permutation of their *orthodox* order $M_1, M_2, \ldots, M_k$ in the macronuclear gene; moreover some MDSs may be inverted in the micronuclear gene. Note also that the IESs from the micronuclear gene are removed (excised) during the gene assembly process.

It has been postulated in [4] and [9] that gene assembly is accomplished through the three molecular operations: *ld-excision*, *hi-excision/reinsertion*, and *dlad-excision/reinsertion*. Each of these operations is of the *fold and recombine* style: first a molecule is folded (and aligned on the pointers), then a cut is made (or cuts are made), and then homologous recombination takes place. The reader is referred to [2] and [1] for more details – but just to have some intuition underlying this paper, we give in Fig. 3, Fig. 4, and Fig. 5 examples for the above three operations.

These three molecular operations provide an *intramolecular* framework for gene assembly: in each operation the molecule involved reacts with itself (and not with another molecule).

Note that folding brings together the specific pointers (involved in a given operation), and as the result of an operation 'smaller' MDSs are spliced together on pointers forming bigger, composite MDSs, and the corresponding IESs are either spliced together forming composite IESs or they are excised. This recombination on pointers creates a more homogenous situation: while, before a recombination took place, there were 'heterogenous' MDS/IES junc-

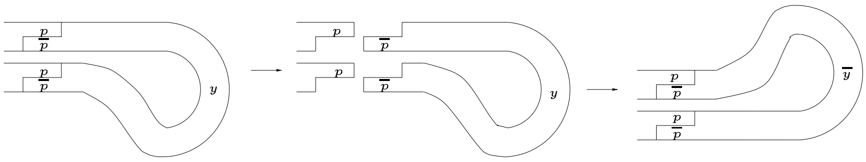


Figure 4: The hi-excision/reinsertion operation.

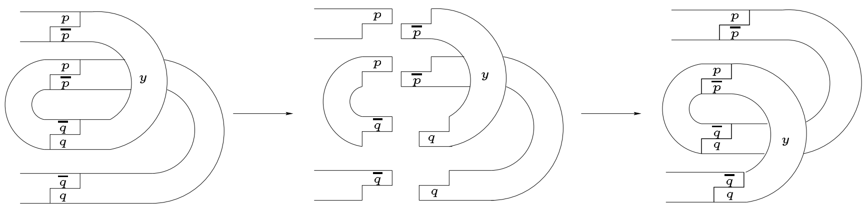


Figure 5: The dlad-excision/reinsertion operation.

tions on involved pointers, after the recombination took place one gets 'homogenous' junctions MDS/MDS and IES/IES (on former pointers). Note that a given sequence of nucleotides forming the two occurrences of the pointer (involved in a given operation) before the recombination, ceases to be a pointer after the recombination, because it ends up either inside a composite MDS or inside a composite IES (recall that a pointer is always located at the boundary between an MDS and an IES).

The above observations underlie our model of gene assembly in ciliates through (cyclic decompositions of) recombination graphs. As a matter of fact our model is more general, and it can be seen as a graph theoretic formulation of the 'fold and recombine' computing paradigm.

First of all, we consider circular graphs, rather than linear ones, to represent also linear DNA molecules. This is technically convenient and does not restrict the generality of our considerations, because a linear graph can be easily closed to a circular graph using one additional vertex, which can be removed whenever we want to restore the linearity. To reflect the MDS/IES structure of a micronuclear (and an intermediate) gene, we consider bicoloured graphs. Besides colouring we use also the labelling function which gives the sequences of nucleotides comprising various (IES or MDS) segments of a DNA molecule.

Hence our general initial situation is a set of circular DNA molecules represented by a bicoloured and labelled graph consisting of circular graphs. The 'fold and recombine' processing of the set of circular molecules is reflected by the two-stage processing of our graphs: first by folding on vertices representing pointers (this is our $*$ operation from Section 4.1), and then unfolding using a pairing function on 'pointer vertices' (this is our $\diamond$ operation from Section 4.2).

In this setup the process of gene assembly becomes the process of dynamic cyclic decomposition of recombination graphs. The successive stages of gene assembly become the successive changes in cyclic decomposition of recombination graphs (on a given set of vertices). Then, the final graph is a set of bicoloured cyclic graphs with genes represented as maximal monochromatic paths.

The paper is organized as follows. In Sections 2 and 3 we recall some basic notions concerning strings and graphs, establishing in this way the notation to be used in this paper. Also, in Section 3 we point out a connection to the work of Pevzner ([6], [7]).

In Section 4 we introduce our tools: the operations of *folding* (a graph), and *unfolding* (a paired graph). We also establish some basic properties of these operations.

In Section 5 we formalize the notions of a *genome* and a *gene*, and study the process of gene assembly by investigating the formal notions of an *assembled genome* and an *assembly strategy*. We prove that for each genome, there exists an assembly strategy which is intracyclic and such that no more than two pointers are needed in each assembly step. We end Section 5 by demonstrating how our general framework applies to gene assembly in ciliates, by modelling an assembly of actin I gene in *Oxytricha nova*.

2 Preliminaries

2.1 Complementary alphabets and strings

If X is a finite set, then $|X|$ denotes the number of its elements. The family of all 2-element subsets of X is denoted by $\mathcal{P}_2(X)$, *i.e.*,

$$\mathcal{P}_2(X) = \{\{x, y\} \mid x, y \in X,\ x \neq y\}.$$

For integers k and n with $k \leq n$, we use $[k, n]$ to denote the interval of the integers between k and n, *i.e.*, $[k, n] = \{k, k+1, \dots, n\}$.

If $\alpha\colon X \to Y$ is a function from a set X to a set Y, and $A \subseteq X$ is a subset of X, then $\alpha \restriction A\colon A \to Y$ is the *restriction* of α to A.

Let $\beta\colon X \to X$ be a permutation of a finite set X. An element $x \in X$ is called a *fixed point* of β, if $\beta(x) = x$. Also, β is an *involution*, if it is of order two, that is, if $\beta(\beta(x)) = x$ for all $x \in X$.

Let Δ be an *alphabet*, that is, a finite set of symbols. The sequences $a_1 a_2 \dots a_n$ with $a_i \in \Delta$ are *strings* (*over* Δ). We use Δ^* to denote the set of all strings over Δ, including the empty string Λ. The alphabet Δ together with an involution $\beta\colon \Delta \to \Delta$ is called a *complementary alphabet*. Such a function β generalizes the Watson - Crick complementarity relation A – T, C – G which holds in the alphabet $\{A, C, G, T\}$ of nucleotides; in this case $\beta(A) = T$, $\beta(T) = A$, and $\beta(C) = G$, $\beta(G) = C$. Note however, that in general we may have for an $a \in \Delta$, $\beta(a) = a$; this does not hold in the alphabet of nucleotides.

The alphabets we consider in this paper are complementary, *i.e.*, they are of the form (Δ, β); also, for each $a \in \Delta$, $\beta(a)$ is denoted simply by $\overline{a}$.

For a complementary alphabet Δ (with an involution $a \mapsto \overline{a}$), we have $\overline{\overline{a}} = a$ for all $a \in \Delta$. We generalize the involution to all strings over Δ by defining the *inversion* in Δ^* as follows: for all $w = a_1 a_2 \dots a_n$ with $a_i \in \Delta$,

$$\overline{w} = \overline{a}_n \overline{a}_{n-1} \dots \overline{a}_1.$$

If we consider the alphabet $\Delta = \{\binom{A}{T}, \binom{T}{A}, \binom{C}{G}, \binom{G}{C}\}$ of the double stranded DNA molecules such that each nucleotide has the complementary nucleotide in the other strand, then

$$\overline{\binom{A}{T}} = \binom{T}{A}, \ \overline{\binom{T}{A}} = \binom{A}{T}, \ \overline{\binom{G}{C}} = \binom{C}{G}, \text{ and } \overline{\binom{C}{G}} = \binom{G}{C},$$

and so, *e.g.*,

$$\overline{\binom{A}{T}\binom{G}{C}\binom{T}{A}\binom{C}{G}} = \binom{G}{C}\binom{A}{T}\binom{C}{G}\binom{T}{A}.$$

2.2 Graphs with labels and colours

We consider graphs with multiple edges and loops together with edge colouring and labelling functions. Each (undirected) edge of a graph is oriented in the opposite directions. A general graph, as defined below, describes a folded DNA molecule divided into MDS and IES regions. The labelling function of a graph attaches a string and its inversion to the two orientations of the edges; such a label represents a submolecule of the DNA molecule represented by the graph. In the case of ciliates, the colour of an edge in the graph represents either an MDS or an IES region: the colours are 1 and 2, where 1 denotes an MDS region and 2 denotes an IES region. However, in our general model we allow arbitrarily many colours $c \in [1, k]$.

To be more precise, let V be a finite set. For each pair $a = (x, y) \in V \times V$, let $\overline{a} = (y, x)$ be its *reverse pair*. Clearly, the mapping $a \mapsto \overline{a}$ is an involution having the fixed points (x, x) for $x \in V$.

A *graph* $\gamma = (V, E, \varepsilon, f, h)$ consists of

- a finite set of *vertices* V, and a finite set of *edges* E together with an involution $e \mapsto \overline{e}$ such that, for all $e \in E$, $e \neq \overline{e}$, and if $e \in E$, then also $\overline{e} \in E$,
- an *end point mapping* $\varepsilon \colon E \to V \times V$ such that, for all $e \in E$, $\varepsilon(\overline{e}) = \overline{\varepsilon(e)}$,

- a *labelling function* $f\colon E \to \Delta^*$, for a (complementary) alphabet Δ, such that for all $e \in E$, $f(\overline{e}) = \overline{f(e)}$,

- a *colouring function* $h\colon E \to [1,k]$, for some $k \geq 1$, such that for all $e \in E$, $h(e) = h(\overline{e})$.

The values of h are called the *colours* of the graph, and γ is said to be *bicoloured*, if it has two colours, *i.e.*, $h\colon E \to \{1,2\}$.

For an edge $e \in E$ with $\varepsilon(e) = (x,y)$, x is the *initial vertex* of e, denoted by $\iota(e)$, and y is the *terminal vertex* of e, denoted by $\tau(e)$; hence $\varepsilon(e) = (\iota(e), \tau(e))$. We also say that $\iota(e)$ and $\tau(e)$ are the *ends* of e. An edge $e \in E$ with $\iota(e) = \tau(e)$ is a *loop*; note that for a loop e, $\varepsilon(e) = \varepsilon(\overline{e})$ (but always $e \neq \overline{e}$). In general, we say that two edges $e, e' \in E$ are *parallel*, if $\varepsilon(e) = \varepsilon(e')$.

If needed, the components of a graph are identified by subscripts, *i.e.*, $\gamma = (V_\gamma, E_\gamma, \varepsilon_\gamma, f_\gamma, h_\gamma)$.

Remark 1. *In order to simplify the notations we shall usually write* $e = (x,y)$ *instead of* $\varepsilon_\gamma(e) = (x,y)$. Note however that this notation can be ambiguous when the graph γ has parallel edges. □

Let $x \in V_\gamma$ be a vertex of γ, and let

$$E^+_\gamma(x) = \{e \in E_\gamma \mid \tau(e) = x\} \quad \text{and} \quad E^-_\gamma(x) = \{e \in E_\gamma \mid \iota(e) = x\}.$$

The *valency* $\mathrm{val}_\gamma(x)$ of x is the number of edges entering x, that is,

$$\mathrm{val}_\gamma(x) = |E^+_\gamma(x)|.$$

Notice that if $e = (x,x)$ is a loop, then both e and $\overline{e}$ are in $E^+_\gamma(x)$ (and in $E^-_\gamma(x)$). Clearly, $|E^-_\gamma(x)| = |E^+_\gamma(x)|$, and so also $\mathrm{val}_\gamma(x) = |E^-_\gamma(x)|$. For each vertex $x \in V_\gamma$ and each colour $c \in [1,k]$, let

$$\mathrm{val}_\gamma(x,c) = |\{e \in E^+_\gamma(x) \mid h_\gamma(e) = c\}|.$$

Hence $\mathrm{val}_\gamma(x,c)$ is the number of edges coloured by c entering x. Clearly, we have $\mathrm{val}_\gamma(x) = \sum_{c=1}^{k} \mathrm{val}_\gamma(x,c)$. A vertex $x \in V_\gamma$ is said to be *balanced*, if for all $c \in [1,k]$,

$$\mathrm{val}_\gamma(x,c) \leq \mathrm{val}_\gamma(x)/2.$$

The graph γ is *balanced*, if every vertex $x \in V_\gamma$ is balanced.

Remark 2. For a bicoloured graph γ, we shall draw the edges e coloured by 1 using thick arrows and those coloured by 2 using thin arrows. Also, the reverse edges of the drawn edges will be omitted from the figures, since their colours and labels are determined by the drawn edges. Note that the choice of the drawn edge, $e = (x,y)$ or $\overline{e} = (y,x)$, is arbitrary. As usual, the label $f_\gamma(e)$ of a drawn edge e is given on the drawing of e. □

Example 2.1. Let $\Delta = \{a, b\}$ be the label alphabet with $\overline{a} = b$. The bicoloured graph γ from Fig. 6 has the vertex set $V_\gamma = [1, 8]$, and it has 24 (oriented) edges. Here $h_\gamma(e) = 1$ for the thick edges (and their reversals), and, *e.g.*, $f_\gamma(e) = abb$ and $f_\gamma(\overline{e}) = aab$ for the edge $e = (7, 8)$. The graph γ is balanced; *e.g.*, $\text{val}_\gamma(1) = 4$ (although only one edge entering 1 is drawn in the figure), and $\text{val}(1, 1) = 2 = \text{val}(1, 2)$. $\square$

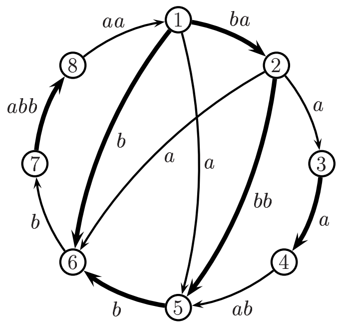


Figure 6: The graph of Example 2.1

A *walk* in a graph γ is a string $\pi = e_1 e_2 \dots e_n$ over E_γ such that $\tau(e_i) = \iota(e_{i+1})$ for $i \in [1, n-1]$. The vertex $\iota(e_1)$ ($\tau(e_n)$, respectively) is the *initial vertex* of π, denoted by $\iota(\pi)$, (the *terminal vertex* of π, respectively, denoted by $\tau(\pi)$). Then, the initial and the terminal vertices of π are called the *ends* of π, and we write $\pi \colon \iota(\pi) \to \tau(\pi)$, or with some ambiguity (for graphs with parallel edges) we present such a walk in a more readable form as a sequence

$$\pi \colon x_1 \to x_2 \to \cdots \to x_{n+1} \quad \text{or} \quad \pi \colon x_1 \xrightarrow{c_1} x_2 \xrightarrow{c_2} \dots \xrightarrow{c_n} x_{n+1},$$

where $e_i = (x_i, x_{i+1})$ and $c_i = h_\gamma(e_i)$. We say that each vertex x_i, for $1 \le i \le n+1$, and each edge e_i, for $1 \le 1 \le n$, is *on the walk* π. The walk π is *closed*, if $\iota(\pi) = \tau(\pi)$. The *label* of the walk π is the string

$$f_\gamma(\pi) = f_\gamma(e_1) f_\gamma(e_2) \dots f_\gamma(e_n).$$

If, for some colour c, $h_\gamma(e_i) = c$ for all $i \in [1, n]$, then the walk π is *monochromatic*, and we write $h_\gamma(\pi) = c$.

For a walk $\pi = e_1 e_2 \dots e_n$, let $\overline{\pi} = \overline{e}_n \overline{e}_{n-1} \dots \overline{e}_1$ be its *reverse walk*. Clearly, the reversed walk is a walk that satisfies

$$f_\gamma(\overline{\pi}) = \overline{f_\gamma(\pi)}.$$

If π_1 and π_2 are walks such that $\tau(\pi_1) = \iota(\pi_2)$, then $\pi_1 \pi_2 \colon \iota(\pi_1) \to \tau(\pi_2)$ is their *composed walk*, where $\pi_1 \pi_2$ is the concatenation of these walks. Clearly, $f_\gamma(\pi_1 \pi_2) = f_\gamma(\pi_1) f_\gamma(\pi_2)$.

Let $\pi = e_1 e_2 \dots e_n$ be a walk such that $e_i = (x_i, x_{i+1})$ for each i. Then π is a *path*, if $x_i \ne x_j$ for all $i \ne j$, and π is a *cycle*, if it is closed and $x_i \ne x_j$ for all $i \ne j$ with $i, j \le n$.

Example 2.2. Let γ be the bicoloured graph of Example 2.1, see Fig. 6. The closed walk $\pi\colon 1 \to 2 \to 5 \to 6 \to 1$ is monochromatic, and $f_\gamma(\pi) = babbb\overline{b} = babbba$. (Notice that the reverse of the edge $e = (6,1)$ is drawn in the graph, and therefore $f_\gamma(e) = a$.) □

A graph γ is said to be *connected*, if for any two vertices $x, y \in V_\gamma$, there exists a path $x \to y$, and γ is *cyclic*, if all its vertices and edges are on one cycle. In this case we assume that γ is given as a cycle.

A graph γ' is a *subgraph* of γ, if $V_{\gamma'} \subseteq V_\gamma$, $E_{\gamma'} \subseteq E_\gamma$, $f_{\gamma'} = f_\gamma \restriction E_{\gamma'}$, and $h_{\gamma'} = h_\gamma \restriction E_{\gamma'}$. If, moreover, $E_{\gamma'} = E_\gamma \cap \varepsilon_\gamma^{-1}(V_{\gamma'} \times V_{\gamma'})$, then γ' is an *induced subgraph* (or *subgraph induced by* the set $V_{\gamma'}$), denoted by $\gamma \restriction V_{\gamma'}$.

For the graphs $\gamma_1, \dots, \gamma_m$ such that $V_{\gamma_i} \cap V_{\gamma_j} = \emptyset$ for all $i \neq j$, their *disjoint union*, denoted by $\gamma = \sum_{i=1}^m \gamma_i$, is the graph with $V_\gamma = \cup_{i=1}^m V_{\gamma_i}$ and $E_\gamma = \cup_{i=1}^m E_{\gamma_i}$ such that $\gamma \restriction V_{\gamma_i} = \gamma_i$.

An (induced) subgraph $\gamma \restriction A$ is a *connected component* of γ, if $\gamma|A$ is a maximal connected subgraph (maximal with respect to vertices and edges). Clearly, each graph γ is partitioned by its connected components $\gamma_i = \gamma \restriction A_i$, for $i \in [1, m]$ and $m \geq 1$, meaning that $\gamma = \sum_{i=1}^m \gamma_i$ and $V_\gamma = \cup_{i=1}^m A_i$.

Two graphs γ and γ' are *isomorphic*, if there are bijections $\alpha\colon V_\gamma \to V_{\gamma'}$ and $\beta\colon E_\gamma \to E_{\gamma'}$ such that for all edges $e \in E_\gamma$ with $\varepsilon_\gamma(e) = (x, y)$, $\varepsilon_{\gamma'}(\beta(e)) = (\alpha(x), \alpha(y))$, $f_{\gamma'}(\beta(e)) = f_\gamma(e)$, and $g_{\gamma'}(\beta(e)) = g_\gamma(e)$.

Remark 3. Note that the notion of isomorphism requires that the corresponding edges, e and $\beta(e)$, have *identical* labels and colours.

In this paper *we shall identify isomorphic graphs*: if the graphs γ and γ' are isomorphic, then we write $\gamma = \gamma'$. □

3 Eulerian graphs

3.1 Alternating trails

Let $\pi = e_1 e_2 \dots e_n$ be a walk in a graph γ. Then π is called

- a *trail*, if $e_i \neq e_j$ and $e_i \neq \overline{e}_j$ for all $i \neq j$,
- a *maximal trail*, if it is a trail, and there exists no edge $e \in E_\gamma$ such that $e_1 e_2 \dots e_n e$ is a trail; we also say then that π is a *maximal trail of* e_1,
- an *Euler trail*, if it is a trail and, for each $e \in E_\gamma$, there is an index i such that either $e = e_i$ or $\overline{e} = e_i$,
- an *alternating* walk, if $h_\gamma(e_i) \neq h_\gamma(e_{i+1})$ for all $i \in [1, n-1]$,
- an *alternating closed walk*, if π is alternating, closed, and $h_\gamma(e_n) \neq h_\gamma(e_1)$.

A graph γ is said to be (*alternating*) *Eulerian*, if γ has an (alternating) closed Euler trail, and γ is *even*, if the valency $\mathrm{val}_\gamma(x)$ is even for each $x \in V_\gamma$.

The following result is known as *Euler's theorem*, see, *e.g.*, West [11].

Theorem 3.1. *A graph γ is Eulerian if and only if it is connected and even.*

Let γ be a graph, and denote by $\gamma - e$ the subgraph of γ with edges $E_\gamma \setminus \{e, \overline{e}\}$. An edge $e \in E_\gamma$ is a *bridge*, if the number of the connected components of $\gamma - e$ is greater than that of γ; otherwise e is a *nonbridge*.

The following procedure for obtaining Euler trails is known as *Fleury's algorithm* (see [11]). Note that the algorithm is nondeterministic, *i.e.*, at each intermediate step there may be several choices how to continue.

Fleury's algorithm: *Let γ be a connected and even graph. Choose any edge $e_1 \in E_\gamma$, and let $\gamma_1 = \gamma - e_1$. Repeat the following for $i \geq 1$:*
(1) if $E^-_{\gamma_i}(\tau(e_i)) = \emptyset$, then the result is $\pi = e_1e_2\ldots e_i$.
(2) if $E^-_{\gamma_i}(\tau(e_i))$ has a nonbridge, then choose one of them to be e_{i+1}; otherwise choose any edge $e_{i+1} \in E^-_{\gamma_i}(\tau(e_i))$. Set $\gamma_{i+1} = \gamma_i - e_{i+1}$. □

Theorem 3.2. *Let γ be a connected even graph. Then the result of Fleury's algorithm is a closed Euler trail of γ, and every closed Euler trail can be thus obtained.*

The following result was proved by Kotzig [5], see also Pevzner [7].

Theorem 3.3 (Kotzig). *An even graph γ is alternating Eulerian if and only if it is connected and balanced.*

In particular, if the graph γ is bicoloured, then it is balanced if and only if $\mathrm{val}_\gamma(x, 1) = \mathrm{val}_\gamma(x, 2)$ for all $x \in V_\gamma$. Therefore, by Theorem 3.3,

Theorem 3.4. *A bicoloured even graph γ is alternating Eulerian if and only if γ is connected and* $\mathrm{val}_\gamma(x, 1) = \mathrm{val}_\gamma(x, 2)$ *for all $x \in V_\gamma$.*

A graph γ is a *recombination graph*, if it is bicoloured, and for each vertex x, $\mathrm{val}_\gamma(x) = 2$ or 4, and every vertex of valency 4 is balanced (*i.e.*, $\mathrm{val}_\gamma(x, 1) = 2 = \mathrm{val}_\gamma(x, 2)$).

Example 3.5. The graph γ of Fig. 6 is a recombination graph that is balanced. An alternating closed Euler trail of γ can be traced using Fleury's algorithm. Starting with the edge $e = (1, 2)$, we obtain the following alternating closed Euler trail π: $1 \xrightarrow{1} 2 \xrightarrow{2} 3 \xrightarrow{1} 4 \xrightarrow{2} 5 \xrightarrow{1} 6 \xrightarrow{2} 7 \xrightarrow{1} 8 \xrightarrow{2} 1 \xrightarrow{1} 6 \xrightarrow{2} 2 \xrightarrow{1} 5 \xrightarrow{2} 1$. □

Recombination graphs are not necessarily alternating Eulerian, since the vertices of valency 2 need not be balanced. However, by the following lemma, they have a closed Euler trail that *alternates at valency* 4 *vertices*, *i.e.*, $\pi = e_1e_2\ldots e_n$ with $e_i = (x_i, x_{i+1})$ and $x_{n+1} = x_1$ such that for all i, $h_\gamma(e_i) \neq h_\gamma(e_{i+1})$ if $\mathrm{val}_\gamma(x_{i+1}) = 4$.

Lemma 3.6. *A connected recombination graph γ has a closed Euler trail π that alternates at valency 4 vertices.*

Proof. If γ is a cyclic graph, then the claim is trivial, since, in this case, all vertices have valency 2. Assume then that γ is not a cyclic graph. Recall that the unbalanced vertices $x \in V_\gamma$ have valency 2, and the two edges in $E_\gamma^+(x)$ have the same colour.

Define a new bicoloured graph γ' as follows: for each unbalanced vertex $x \in V_\gamma$ with $\mathrm{val}_\gamma(x, c) = 2$ (where either $c = 1$ or $c = 2$), add a loop $e = (x, x)$ with $h_{\gamma'}(e) = 3 - c$ and $f_{\gamma'}(e) = \Lambda$. Obviously, the graph γ' is balanced, and therefore it has an alternating closed Euler trail π', by Theorem 3.3. Clearly, the corresponding path π in the original recombination graph γ, where the introduced loops are removed, satisfies the claim. $\square$

For the bicoloured graphs, Pevzner [6] proved that any two alternating closed Euler trails can be obtained from each other by using relatively simple transformations of closed walks. The *exchange operation* transforms a closed walk with a decomposition $\pi = \pi_1\pi_2\pi_3\pi_4\pi_5$, where $\iota(\pi_2) = \iota(\pi_4)$ and $\tau(\pi_2) = \tau(\pi_4)$, to the closed walk $\pi' = \pi_1\pi_4\pi_3\pi_2\pi_5$. The *reflection operation* transforms a closed walk with a decomposition $\pi = \pi_1\pi_2\pi_3$, where π_2 is a closed walk, to the closed walk $\pi' = \pi_1\overline{\pi}_2\pi_3$. It was proved by Pevzner that if γ is a bicoloured graph, then every two alternating closed Euler trails can be transformed to each other by a finite number of exchange and reflection operations that preserve alternating closed walks.

3.2 Pairing functions of even graphs

Let γ be an even graph. A *pairing* ψ_x of a vertex $x \in V_\gamma$ is a bijective mapping $\psi_x \colon E_\gamma^+(x) \to E_\gamma^-(x)$ that respects inversions, *i.e.*, ψ_x satisfies the conditions:

$$\psi_x(\overline{\psi_x(e)}) = \overline{e}, \tag{1}$$

$$\psi_x(e) = \overline{e} \iff e \text{ is a loop.} \tag{2}$$

By (1), if $\psi_x(e_1) = e_2$ for $e_1 \in E_\gamma^+(x)$, then $\psi_x(\overline{e}_2) = \overline{e}_1$. A *pairing function* of γ is a function $\psi \colon x \mapsto \psi_x$ of the vertices such that ψ_x is a pairing for each $x \in V_\gamma$.

Let ψ be a pairing function of an even graph γ. For each edge $e_1 \in E_\gamma$, let $\pi_\psi(e_1) = e_1e_2\dots e_k$ be the maximal trail of e_1 such that $\psi_{x_i}(e_{i-1}) = e_i$ for all $2 \le i \le k$, where $e_i = (x_i, x_{i+1})$. Clearly, $\pi_\psi(e)$ is well defined for each $e \in E_\gamma$, since ψ is a bijection. Note also that if $e \in E_\gamma$ is a loop, then $\pi_\psi(e) = e$.

For a recombination graph γ, the *natural pairing function* $\psi \colon x \mapsto \psi_x$ is defined by requiring that ψ_x pairs the same coloured edges for vertices of valency 4, and the only possible edges for the vertices x of valency 2. Formally, if $e \in E_\gamma^+(x)$, $e' \in E_\gamma^-(x)$ with $\overline{e} \neq e'$, then

$$\psi_x(e) = e' \iff \text{ either } \mathrm{val}_\gamma(x) = 2, \text{ or } \mathrm{val}_\gamma(x) = 4 \text{ and } h_\gamma(e) = h_\gamma(e').$$

Example 3.7. Consider the recombination graph γ from Fig. 6, and let $\psi\colon x \to \psi_x$ be the natural pairing function of γ. Then, for instance, $\psi_1(8,1) = (1,5)$ (and $\psi_1(5,1) = (1,8)$), $\psi_1(2,1) = (1,6)$ (and $\psi_1(6,1) = (1,2)$), and $\psi_3(2,3) = (3,4)$ (and $\psi_3(4,3) = (3,2)$). The pairing function ψ gives a partition of the edges of γ into the following two closed trails (and their reversals): $\pi_1\colon 1 \xrightarrow{1} 2 \xrightarrow{1} 5 \xrightarrow{1} 6 \xrightarrow{1} 1$ and $\pi_2\colon 1 \xrightarrow{2} 5 \xrightarrow{2} 4 \xrightarrow{1} 3 \xrightarrow{2} 2 \xrightarrow{2} 6 \xrightarrow{2} 7 \xrightarrow{1} 8 \xrightarrow{2} 1$. □

The following result was proved by Tucker [10].

Theorem 3.8 (Tucker). *Let ψ be a pairing function of an even graph γ. Then the maximal trail $\pi_\psi(e)$ is closed for each $e \in E_\gamma$, and the edge sets of the maximal trails $\pi_\psi(e)$, $e \in E_\gamma$, form a partition of E_γ.*

4 Operations on graphs

4.1 Folding a graph

In the graph theoretical framework of this paper, the process of gene assembly in ciliates will be divided into two stages: first, folding the circular graphs (corresponding to circular DNA molecules), and, second, unfolding the folded graphs by splitting the vertices of valency 4. In this section, we take a somewhat more general point of view.

A pair $p = \{x, y\} \in \mathcal{P}_2(V_\gamma)$ of different vertices of a graph γ is called a *pointer*, and the vertices x and y are called the *ends* of p. A set P of mutually disjoint pointers is called a *pointer set.*

Let $p \in \mathcal{P}_2(V_\gamma)$ be a pointer of γ. Let $V = (V_\gamma \setminus p) \cup \{p\}$, where we assume that p is a new vertex. Let $\varphi_p\colon V_\gamma \to V$ be a mapping defined by

$$\varphi_p(z) = \begin{cases} z, & \text{if } z \notin p, \\ p, & \text{if } z \in p. \end{cases}$$

The *p-folded graph* of γ is the graph

$$\gamma * p = (V, E_\gamma, \varepsilon, f_\gamma, h_\gamma)$$

obtained by identifying (*i.e.*, by contracting) the ends of p. More formally, for each edge $e \in E_\gamma$ with $\varepsilon_\gamma(e) = (u, v)$,

$$\varepsilon(e) = (\varphi_p(u), \varphi_p(v)).$$

Note that the folded graph $\gamma * p$ has the same set of edges, and the same labelling and colouring functions as the original graph γ. However, $|V_{\gamma * p}| = |V_\gamma| - 1$, and some edges may become parallel in $\gamma * p$ even though they are not parallel in γ. Indeed, if $p = \{x, y\}$, $\varepsilon_\gamma(e_1) = (x, z) \in E_\gamma$ and $\varepsilon_\gamma(e_2) = (y, z) \in E_\gamma$, then $\varepsilon_{\gamma * p}(e_1) = (p, z) = \varepsilon_{\gamma * p}(e_2) \in E_{\gamma * p}$.

Example 4.1. Let γ be the recombination graph from Fig. 6, and let $p = \{1, 7\}$. Then the graph $\gamma * p$ is given in Fig. 7. Note that $\gamma * p$ is not a recombination graph, since the valency of p is six. However, $\gamma * p$ is still balanced. $\square$

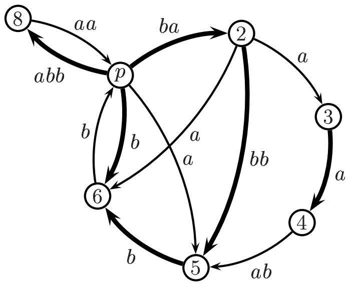


Figure 7: The p-folded graph of the graph of Fig. 6

If $p_1 \in \mathcal{P}_2(V_\gamma)$ and $p_2 \in \mathcal{P}_2(V_{\gamma * p_1})$ are pointers of the graphs γ and $\gamma * p_1$, respectively, then we define $\gamma * p_1 * p_2 = (\gamma * p_1) * p_2$, and, inductively,

$$\gamma * p_1 * \ldots * p_m = (\gamma * p_1 * \ldots * p_{m-1}) * p_m$$

for pointers p_i of $\gamma * p_1 * p_2 * \ldots * p_{i-1}$.

Lemma 4.2. *Let γ be a balanced graph, and let $p_1, \ldots, p_m$ be a sequence such that p_1 is a pointer of γ and p_i is a pointer of $\gamma * p_1 * p_2 * \ldots * p_{i-1}$ for $i = 2, \ldots, m$. Then the graph $\gamma * p_1 * p_2 * \ldots * p_m$ is balanced.*

Proof. Assume that γ has k colours. For a pointer $p = \{x, y\}$ and a colour $c \in [1, k]$, we have, by the construction of $\gamma * p$,

$$\mathrm{val}_{\gamma * p}(z, c) = \begin{cases} \mathrm{val}_\gamma(z, c) & \text{if } z \neq p, \\ \mathrm{val}_\gamma(x, c) + \mathrm{val}_\gamma(y, c) & \text{if } z = p. \end{cases}$$

In particular, $\mathrm{val}_{\gamma * p}(z) = \mathrm{val}_\gamma(z)$ for all $z \notin p$, and, in this case, $\mathrm{val}_\gamma(z, c) \leq (1/2)\,\mathrm{val}_{\gamma * p}(z) = (1/2)\,\mathrm{val}_\gamma(z)$. Moreover, $\mathrm{val}_{\gamma * p}(p) = \mathrm{val}_\gamma(x) + \mathrm{val}_\gamma(y)$, and

$$\begin{aligned} \mathrm{val}_\gamma(p, c) &= \mathrm{val}_\gamma(x, c) + \mathrm{val}_\gamma(y, c) \leq (1/2)\,(\mathrm{val}_\gamma(x) + \mathrm{val}_\gamma(y)) \\ &= (1/2)\,\mathrm{val}_{\gamma * p}(z). \end{aligned}$$

This shows that if γ is balanced, then so is the p-folded graph $\gamma * p$. The claim of the lemma follows inductively from this argument. $\square$

The following lemma is straightforward to prove.

Lemma 4.3. *Let $\{p,q\}$ be a pointer set of a graph γ. Then $\gamma * p * q = \gamma * q * p$.*

For a pointer set P of a graph γ, we define the *P-folded graph* $\gamma * P$ as

$$\gamma * P = \gamma * p_1 * \ldots * p_m,$$

where $p_1, \ldots, p_m$ is an arbitrary permutation of P. By Lemma 4.3, $\gamma * P$ is well defined. Also, if $P_1, P_2, \ldots, P_n$ are disjoint pointer sets, then we shall write $\gamma * P_1 * P_2 * \ldots * P_n$ for $(...((\gamma * P_1) * P_2) * \ldots) * P_n$. By Lemma 4.3, we have then

$$\gamma * P_1 * P_2 * \ldots * P_n = \gamma * \bigcup_{i=1}^{n} P_i. \tag{3}$$

4.2 Unfolding paired graphs

Let γ be an even graph together with a pairing function $\psi \colon x \mapsto \psi_x$ of its edges. For a vertex $x \in V_\gamma$, let $(e_{11}, e_{12}, e_{21}, e_{22}, \ldots, e_{m1}, e_{m2})$ be an ordering of $E_\gamma^+(x)$, where $\psi_x(e_{i1}) = \overline{e}_{i2}$ (and hence $\psi_x(e_{i2}) = \overline{e}_{i1}$). Let $x^1, x^2, \ldots, x^m$ be new vertices. The *ψ-unfolded graph γ at x* is the graph

$$\gamma \diamond_\psi x = (V, E_\gamma, \varepsilon, f_\gamma, h_\gamma),$$

such that $V = (V_\gamma \setminus \{x\}) \cup \{x^1, \ldots, x^m\}$ and for each $e \in E_\gamma$, $\varepsilon(e) = \varepsilon_\gamma(e)$, if x is not an end of e, and

$$\varepsilon(e_{ij}) = \begin{cases} (y, x^i) & \text{if } \varepsilon_\gamma(e_{ij}) = (y, x) \text{ and } y \neq x, \\ (x^i, x^i) & \text{if } \varepsilon_\gamma(e_{ij}) = (x, x), \end{cases}$$

and $\varepsilon(\overline{e}_{ij}) = \overline{\varepsilon(e_{ij})}$.

Notice that the ψ-unfolded graph has the same set of edges, and the same labelling and colouring functions as γ. Note also that the pairing function ψ remains as a pairing function of $\gamma \diamond_\psi x$.

Example 4.4. Consider again the recombination graph γ of Fig. 6, and let ψ be the natural pairing function of γ. Then $\gamma \diamond_\psi 1$ is given in Fig. 8. □

By the construction of unfolding, we have immediately the following result.

Lemma 4.5. *Let γ be an even graph with a pairing function ψ, and $x \in V_\gamma$. Then $\mathrm{val}_{\gamma \diamond_\psi x}(x^i) = 2$ for each new vertex x^i of $\gamma \diamond_\psi x$, and, $\mathrm{val}_{\gamma \diamond_\psi x}(y) = \mathrm{val}_\gamma(y)$, otherwise. Moreover, if γ is a recombination graph, then so is $\gamma \diamond_\psi x$.*

For different vertices $x_1, x_2 \in V_\gamma$, we write $\gamma \diamond_\psi x_1 \diamond_\psi x_2 = (\gamma \diamond_\psi x_1) \diamond_\psi x_2$, and, inductively, $\gamma \diamond_\psi x_1 \diamond_\psi \ldots \diamond_\psi x_m = (\gamma \diamond_\psi x_1 \diamond_\psi \ldots \diamond_\psi x_{m-1}) \diamond_\psi x_m$ for different vertices $x_i \in V_\gamma$.

The following lemma follows directly from the definition of unfolding.

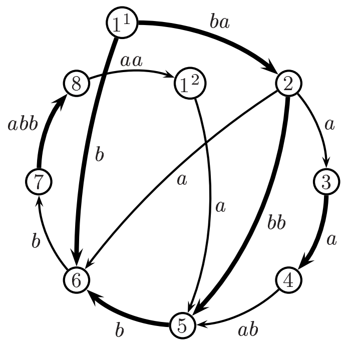


Figure 8: The ψ-unfolded graph $\gamma \diamond_\psi 1$ of Example 4.4

Lemma 4.6. *Let $x, y \in V_\gamma$ be different vertices of an even graph γ with a pairing function ψ. Then $\gamma \diamond_\psi x \diamond_\psi y = \gamma \diamond_\psi y \diamond_\psi x$.*

By Lemma 4.6, for an even graph γ with a pairing function ψ, and a subset $A \subseteq V_\gamma$, we can write

$$\gamma \diamond_\psi A = \gamma \diamond_\psi x_1 \diamond_\psi x_2 \diamond_\psi \ldots \diamond_\psi x_m,$$

where $x_1, x_2, \ldots, x_m$ is any permutation of A. Furthermore, let $\gamma \diamond_\psi A_1 \diamond_\psi A_2 \diamond_\psi \ldots \diamond_\psi A_n = (\ldots((\gamma \diamond_\psi A_1) \diamond_\psi A_2)\ldots) \diamond_\psi A_n$ for disjoint subsets $A_1, A_2, \ldots, A_n \subseteq V_\gamma$. By Lemma 4.6,

$$\gamma \diamond_\psi A_1 \diamond_\psi A_2 \diamond_\psi \ldots \diamond_\psi A_n = \gamma \diamond_\psi \bigcup_{i=1}^{n} A_i. \tag{4}$$

For an even graph γ with a pairing function ψ, let

$$F(\gamma) = \{x \in V_\gamma \mid \mathrm{val}_\gamma(x) \geq 4\}.$$

Then the graph $\gamma \diamond_\psi F(\gamma)$ is called the *ψ-unfolded graph* of γ.

Lemma 4.7. *If γ is an even graph with a pairing function ψ, then its ψ-unfolded graph is a disjoint union of cyclic graphs.*

Proof. For each vertex x of $\gamma \diamond_\psi F(\gamma)$, $\mathrm{val}_{\gamma \diamond_\psi F(\gamma)}(x) = 0$ or 2. This follows by Lemma 4.5 and the fact that each vertex $x \notin F(\gamma)$ satisfies this property. This is equivalent to the claim of the lemma. $\square$

Let γ be a graph with a pointer set P, and let ψ be a pairing function of the P-folded graph $\gamma * P$. We denote

$$\gamma \circledast_\psi P = (\gamma * P) \diamond_\psi P.$$

We shall write $\gamma \circledast P$ for $\gamma \circledast_\psi P$, if $\gamma * P$ is a recombination graph and ψ is its natural pairing function.

Lemma 4.8. *Let γ be a disjoint union of cyclic graphs. Let P be a pointer set of γ, and let ψ be a pairing function of the P-folded graph $\gamma * P$. Then $\gamma \circledast_\psi P$ is a disjoint union of cyclic graphs.*

Proof. By Lemma 4.7, $(\gamma * P) \diamond_\psi F(\gamma * P)$ is a disjoint union of cyclic graphs. The present claim follows from the fact that $F(\gamma * P) = P$. $\square$

We will also need the following simple observation.

Lemma 4.9. *Let γ be a bicoloured graph with a pairing function ψ. Let $A \subseteq V_\gamma$, and let $P = \{\{x^1, x^2\} \mid x \in A\}$. Then $(\gamma \diamond_\psi P) * P = \gamma$.*

Recall, see Remark 3, that we identify isomorphic graphs. In Lemma 4.9 the identifying isomorphism is: $\{x^1, x^2\} \mapsto x$.

5 Genomes and assembled graphs

5.1 Assembled graphs of genomes

In this section we shall study recombination graphs. Recall that these graphs are even, bicoloured, the valencies of the vertices are equal to either 2 or 4, and each vertex of valency 4 is balanced. Also, unless stated otherwise, the pairing function of a recombination graph will be the natural pairing function.

Let γ be a bicoloured cyclic graph with the vertex set $V_\gamma = \{x_1, \ldots, x_n\}$ and the edge set $E_\gamma = \{e_1, \ldots, e_n, \overline{e}_1, \ldots, \overline{e}_n\}$, where $e_i = (x_i, x_{i+1})$ and $x_{n+1} = x_1$. A vertex x_i is a *boundary vertex* of γ, if $h_\gamma(e_{i-1}) \neq h_\gamma(e_i)$, where $i-1$ is modulo n. We denote by $B(\gamma)$ the set of all boundary vertices of γ. A monochromatic path π is a *segment*, if $h_\gamma(\pi) = 1$ and the ends of π are boundary vertices.

Notice that a cyclic graph γ is monochromatic if and only if it has no boundary vertices. Clearly, either each edge e with $h_\gamma(e) = 1$ of a cyclic graph γ belongs to a unique segment, or γ is monochromatic.

For a disjoint union $\gamma = \sum_{i=1}^m \gamma_i$ of bicoloured cyclic graphs γ_i, we let its boundary vertex set be $B(\gamma) = \cup_{i=1}^m B(\gamma_i)$.

A pair $\mathcal{G} = (\gamma, P)$ is a *genome*, if $\gamma = \sum_{i=1}^m \gamma_i$ is a disjoint union of bicoloured cyclic graphs γ_i, for $i \in [1, m]$, and $P \subseteq \mathcal{P}_2(B(\gamma))$ is a pointer set of γ involving boundary vertices only.

Remark 4. The micronuclear DNA molecule is linear. However, our choice to consider circular graphs (corresponding to circular DNA molecules) is not a restriction, since each *linear graph*, *i.e.*, a graph that is a path, can be closed to a cyclic graph as follows. Let γ be a linear graph with the vertex set $V_\gamma = \{x_1, x_2, \ldots, x_n\}$ and the set of edges $E_\gamma = \{e_1, \overline{e}_1, e_2, \overline{e}_2, \ldots, e_{n-1}, \overline{e}_{n-1}\}$, where $e_i = (x_i, x_{i+1})$. Let γ' be the cyclic graph, which is obtained from γ by adding a new vertex x_0, and by adding the edges $\{e_0, \overline{e}_0, e_n, \overline{e}_n\}$ such that $e_0 = (x_0, x_1)$ and $e_n = (x_n, x_0)$, $f_{\gamma'}(e_0) = \Lambda = f_{\gamma'}(e_n)$ and $h_{\gamma'}(e_0) = 2 = h_{\gamma'}(e_n)$. The new vertex x_0 is not a boundary vertex of γ', and therefore it is never identified or split by the operations of folding and unfolding that we will use to assemble genomes, which means that in the assembled genome the vertex x_0 remains to be a nonboundary vertex, and when it is removed from the unfolded graph $\gamma \circledast P$, a linear graph is recovered from the corresponding cyclic graph. □

Example 5.1. The pair (γ, P) with γ given in Fig. 9, and $P = \{p, q\}$, where $p = \{2, 9\}$ and $q = \{5, 8\}$, is a genome. The label alphabet of γ is $\Delta = \{a, b\}$ with $\overline{a} = b$. Note that every vertex of γ is boundary. □

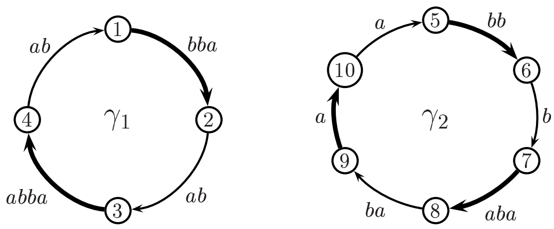


Figure 9: A genome of Example 5.1

Let $\mathcal{G} = (\gamma, P)$ be a genome, and let $R \subseteq P$. The *R-assembled version* of $\mathcal{G}$ is the pair

$$A(\mathcal{G}, R) = (\gamma \circledast R, P \setminus R).$$

Thus the pointers of $P \setminus R$ are not used (they are 'dormant') during the assembly of $A(\mathcal{G}, R)$. The *assembled genome* of $\mathcal{G}$ is the P-assembled version of $\mathcal{G}$, and it is denoted by $A(\mathcal{G})$. Two genomes $\mathcal{G}$ and $\mathcal{G}'$ are *equivalent*, if they have the same assembled genome, $A(\mathcal{G}) = A(\mathcal{G}')$.

Theorem 5.2. *Let $\mathcal{G} = (\gamma, P)$ be a genome, and let $R \subseteq P$. Then $A(\mathcal{G}, R)$ is a genome. In particular, $A(\mathcal{G})$ is a genome.*

Proof. By Lemma 4.8, $\gamma \circledast R$ is a disjoint union of cyclic graphs. □

Let $\mathcal{G} = (\gamma, P)$ be a genome. Each segment of the unfolded graph $\gamma \circledast P$ is an *noncircular gene* of $\mathcal{G}$, and each monochromatic cyclic component of colour 1 of $\gamma \circledast P$ is a *circular gene* of $\mathcal{G}$. Hence, in general, the set of genes of $\mathcal{G}$ consist of the noncircular and the circular genes.

Each gene g of a genome $\mathcal{G}$, which is not a circular gene already, gets assembled from various segments of $\mathcal{G}$, that are called the *parts* of the gene. These parts of g can lie on different cycles of $\mathcal{G}$.

Example 5.3. Let $\mathcal{G}$ be the genome from Example 5.1, see Fig. 9. Here the pointer set is $P = \{p, q\}$, where $p = \{2, 9\}$ and $q = \{5, 8\}$. The P-folded graph $\gamma * P$ is given in Fig. 10. When the unfolding is performed, we obtain the graph $\gamma \circledast P$ in Fig. 11. The genes of $\mathcal{G}$ are all noncircular, and they are $g_1 \colon 1 \to p^1 \to 10$ (with the value $bbaa$), $g_2 \colon 7 \to q^1 \to 6$ (with the value $ababb$), and $g_3 \colon 3 \to 4$ (with the value $abba$). □

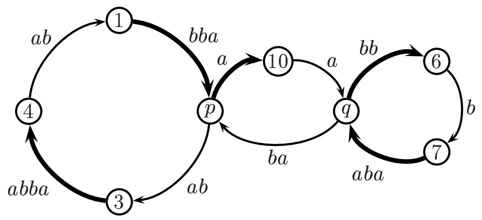


Figure 10: $\gamma * P$ of Example 5.1

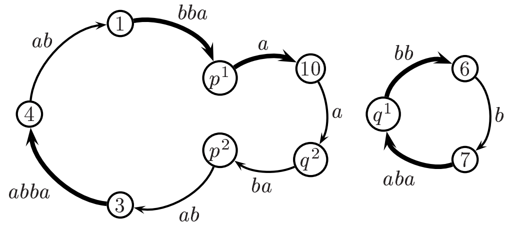


Figure 11: $\gamma \circledast P$ of Example 5.1

5.2 Intracyclic unfolding

For a genome $\mathcal{G} = (\gamma, P)$, the P-folded graph $\gamma * P$ is a recombination graph that has no more connected components than γ. The graph $\gamma * P$ has less connected components than γ only in the case, where there exists a pointer $p \in P$ the ends of which lie in different connected components of γ. We will show now that each recombination graph γ with t connected components can be obtained from a set of t cyclic graphs by folding. From the genome assembly point of view this means that for each genome $\mathcal{G} = (\gamma, P)$ there exists an equivalent genome $\mathcal{G}' = (\gamma', P')$ (*i.e.*, $\gamma * P = \gamma' * P'$) such that each gene of $\mathcal{G}'$ lies on one cyclic component of γ', (*i.e.*, all parts of the gene are on one cyclic graph of γ).

Theorem 5.4. *Let γ be a connected recombination graph. There exists a bicoloured cyclic graph γ' and a set $P \subseteq \mathcal{P}_2(B(\gamma'))$ such that $\gamma' * P = \gamma$.*

Proof. By Lemma 3.6, γ has a closed Euler trail $\pi = e_1 e_2 \dots e_n$ that alternates on valency 4 vertices. Let $e_i = (x_i, x_{i+1})$, where $x_1 = x_{n+1}$, and let $\psi \colon x \mapsto \psi_x$ be the pairing function of γ defined by $\psi_{x_{i+1}}(e_i) = e_{i+1}$ for all $i \in [1, n]$ (where $e_{n+1} = e_1$). Since π is a closed Euler trail, ψ_x is well defined. Now, by Lemma 4.7, $\gamma \diamond_\psi F(\gamma)$ is a disjoint union of cyclic graphs.

By the choice of ψ, $\gamma \diamond_\psi F(\gamma)$ is connected, and since ψ maps each edge to a differently coloured edge, the new vertices x^i for each $x \in F(\gamma)$ are boundary vertices of γ'. The claim follows now from Lemma 4.9, because $(\gamma \diamond_\psi F(\gamma)) * F(\gamma) = \gamma$. □

Example 5.5. Let γ be the recombination graph of Fig. 10. Then $\gamma = \gamma' * P$ for the cyclic graph γ' of Fig. 12, and $P = \{\{p^1, p^2\}, \{q^1, q^2\}\}$. □

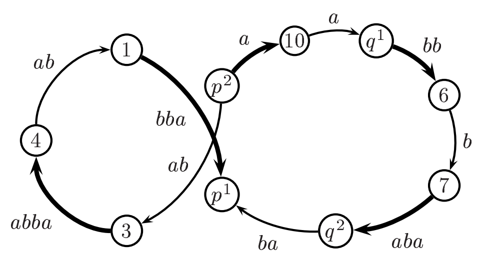


Figure 12: The bicoloured cyclic γ' such that $\gamma = \gamma' * P$

For a genome $\mathcal{G} = (\gamma, P)$, a sequence $\mathcal{S} = (P_1, P_2, \ldots, P_m)$ of subsets of P is an *assembly strategy* of $\mathcal{G}$, if $\{P_1, \ldots, P_m\}$ is a partition of P, *i.e.*, the P_i's are mutually disjoint subsets such that $P = \cup_{i=1}^{m} P_i$.

Lemma 5.6. *Let $\mathcal{G} = (\gamma, P)$ be a genome, $R \subseteq P$ and $p_0 \in P \setminus R$. Then*

$$(\gamma \circledast R) \circledast p_0 = \gamma \circledast (R \cup \{p_0\}).$$

Proof. Let $\gamma_1 = (\gamma \circledast R) \circledast p_0$ and $\gamma_2 = \gamma \circledast (R \cup \{p_0\})$. The graphs γ_1 and γ_2 (as well as all the intermediate graphs between γ and γ_i) have the same set of edges, the same labelling and colouring functions as γ. Therefore it is sufficient to prove that for all $e \in E_{\gamma_1}$, $\varepsilon_{\gamma_1}(e) = \varepsilon_{\gamma_2}(e)$.

Let $R = \{p_1, \ldots, p_m\}$, and let $p_i = \{x_i, y_i\}$ for $i \in [0, m]$. Thus the set of the new vertices of γ_1 and of γ_2 is $A = \{p_i^1, p_i^2 \mid i \in [0, m]\}$.

Let $\varepsilon_{\gamma_1}(e) = (x, y)$ for an edge $e \in E_{\gamma_1}$.

First of all, if $x \neq p_0^r$ and $y \neq p_0^r$ for $r = 1$ and 2, then $\varepsilon_{\gamma \circledast R}(e) = (x, y)$ and thus also $\varepsilon_{\gamma_2}(e) = (x, y)$, because p_0^1 and p_0^2 are not ends of e.

Assume then that $\varepsilon_{\gamma_1}(e) = (p_i^r, p_0^r)$ for some $1 \leq i \leq m$ and $r \in \{1, 2\}$, and so $h_\gamma(e) = r$. Now, $\varepsilon_{(\gamma \circledast R) * p_0}(e) = (p_i^r, p_0)$, $\varepsilon_{\gamma \circledast R}(e) = (p_i^r, z_0)$, where $z_0 \in \{x_0, y_0\}$, and hence $\varepsilon_{\gamma * R}(e) = (p_i, z_0)$. Consequently, $\varepsilon_{\gamma * R * p_0}(e) = (p_i, p_0)$, where $\gamma * R * p_0 = \gamma * (R \cup \{p_0\})$, see (3). Therefore, $\varepsilon_{\gamma \circledast (R \cup \{p_0\})}(e) = (p_i^r, p_0^r)$, since $h_{\gamma * (R \cup \{p_0\})}(e) = r$. Hence, also in this case, $\varepsilon_{\gamma_1}(e) = \varepsilon_{\gamma_2}(e)$.

The case, where $\varepsilon_{\gamma_1}(e) = (x, p_0^r)$ for some $x \in V_\gamma$, is similar, but somewhat easier than the above case. Finally note that if $r \neq s$, there are no edges having the ends p_i^r and p_j^s for any i and j. Hence the lemma holds. □

Lemma 5.6 gives inductively the following generalization.

Lemma 5.7. *Let $\mathcal{G} = (\gamma, P)$ be a genome, and let $P_1, P_2 \subseteq P$ be disjoint. Then $(\gamma \circledast P_1) \circledast P_2 = \gamma \circledast (P_1 \cup P_2) = (\gamma \circledast P_2) \circledast P_1$.*

For disjoint sets $P_1, P_2, \ldots, P_m \subseteq P$ of pointers, Lemma 5.7 allows us to write $\gamma \circledast P_1 \circledast P_2 \circledast \ldots \circledast P_m = (\ldots((\gamma \circledast P_1) \circledast P_2) \circledast \ldots) \circledast P_m$. In particular,

$$\gamma \circledast P_1 \circledast P_2 \circledast \ldots \circledast P_m = \gamma \circledast \bigcup_{i=1}^{m} P_i. \tag{5}$$

Lemma 5.7 together with (5) guarantees that every assembly strategy produces the same assembled genome:

Theorem 5.8. *For every assembly strategy $\mathcal{S} = (P_1, P_2, \ldots, P_m)$ of a genome $\mathcal{G} = (\gamma, p)$, $\gamma \circledast P = \gamma \circledast P_1 \circledast P_2 \circledast \ldots \circledast P_m$.*

In particular, for any genome $\mathcal{G} = (\gamma, P)$ with $P = \{p_1, p_2, \ldots, p_m\}$, the assembly strategy $\mathcal{S} = (\{p_1\}, \{p_2\}, \ldots, \{p_m\})$ consisting of the singleton sets yields the assembled genome $A(\mathcal{G})$: $\gamma \circledast p_1 \circledast \ldots \circledast p_m = \gamma \circledast P$. However, such an assembly strategy can be 'intercircular' in the sense that parts of a gene g of $\mathcal{G}$ that lie in the same connected component of $\gamma \circledast \{p_1, \ldots, p_i\}$ can be in different connected components of $\gamma \circledast \{p_1, \ldots, p_{i+1}\}$ (and they are 'reunited' in the same connected component in the final graph $\gamma \circledast P$). We shall look now for simple assembly strategies for genomes that are 'intracyclic'.

A pointer set $R \subseteq P$ of a genome $\mathcal{G} = (\gamma, P)$ is called *intracyclic*, if for every gene g of $\mathcal{G}$, any two parts g' and g'' of g that lie in the same connected component of γ, lie in the same connected component of $\gamma \circledast R$.

Furthermore, we say that two boundary vertices $x, y \in B(\gamma)$ of a bi-coloured cyclic graph γ are *opposite*, if $\gamma = \pi_1\pi_2\pi_3\pi_4\pi_5$, where π_2 and π_4 are different segments of γ such that $\iota(\pi_2) = x$ and $\tau(\pi_4) = y$, or $\tau(\pi_2) = x$ and $\iota(\pi_4) = y$. If x and y are not opposite, they are *similar*. Note that being similar includes the case where the two vertices are the ends of one segment.

We begin by considering intracyclicity for singleton pointer sets. First, we consider the case when the ends of the pointer are in different connected components of the genome, or they are similar and in the same connected component.

Lemma 5.9. *Let $\mathcal{G} = (\gamma, P)$ be a genome, and $p \in P$ a pointer. Then $\{p\}$ is intracyclic, if the ends of p are (a) in different connected component of γ, or (b) similar in the same connected component of γ.*

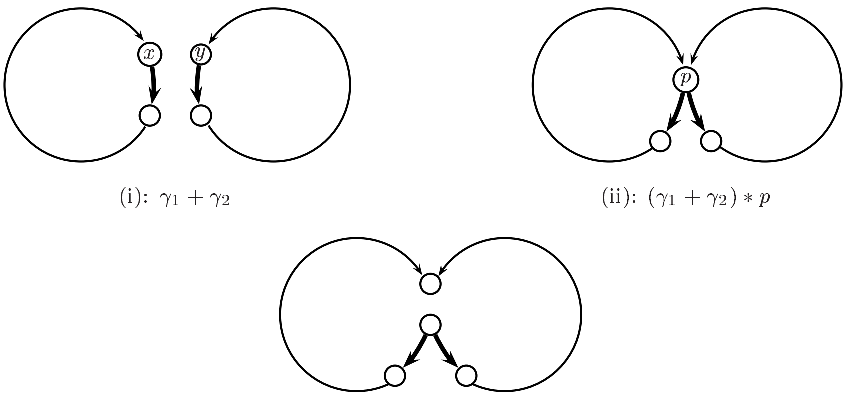
Proof. Let $p = \{x, y\}$. Since p is a pointer, x and y are boundary vertices.

Case (a) is illustrated in Fig. 13. In Fig. 13(i) the connected components γ_1 and γ_2 of γ containing x and y, respectively, are shown together with the segments for which x and y are the initial vertices. Then (ii) gives $(\gamma_1 + \gamma_2) * p$, and (iii) shows the connected component $(\gamma_1 + \gamma_2) \circledast p$ of $\gamma \circledast p$. Since none of the connected components of γ are disconnected to two or more circular graphs during the process, it follows that $\{p\}$ is intracyclic.

In Case (b) we consider two subcases, (b.1) and (b.2), given by Fig. 14 and Fig. 15, respectively.

In Case (b.1), x and y are vertices of different segments as illustrated in Fig. 14(i). Then folding of γ_1 on p yields $\gamma_1 * p$, see Fig. 14(ii), and the unfolding on p yields $\gamma_1 \circledast p$, see Fig. 14(iii). No connected component of γ is disconnected during the process, and hence the claim follows.

In Case (b.2), x and y are the two boundary vertices of the same segment as illustrated in Fig. 15(i). The folding of γ_1 on p yields $\gamma_1 * p$ given in Fig. 15(ii), and then the unfolding on p yields $\gamma_1 \circledast p$ given in Fig. 15(iii). If the segment delimited by x and y in γ_1 is a part of a gene g of the genome, then this part must be the whole g, because it forms a connected component of $\gamma_1 \circledast p$ that has no boundary vertices, and hence will not be combined anymore with any other connected components. Since the connected components of γ other than γ_1 remain as they were, the claim follows in this case also. $\square$



(i): $\gamma_1 + \gamma_2$

(ii): $(\gamma_1 + \gamma_2) * p$

(iii): $(\gamma_1 + \gamma_2) \circledast p$

Figure 13: One pointer cases of Lemma 5.9(a)

Lemma 5.10. *Let $\mathcal{G} = (\gamma, P)$ be a genome. Let $p = \{x, y\} \in P$ be a pointer, where x, y are opposite and they lie on the same cyclic graph γ_i of γ. Then either $\{p\}$ is intracyclic or there exists a pointer q such that $\{p, q\}$ is intracyclic.*

Proof. In transforming γ to $\gamma \circledast p$, the cyclic subgraph γ_i is disconnected into two cyclic graphs γ' and γ'' of $\gamma \circledast p$, see Fig. 16. If $\{p\}$ is not intracyclic for $\mathcal{G}$, then γ' and γ'' contain each a part of a gene g of $\mathcal{G}$, and since g is a path in $\gamma \circledast P$, there exists a pointer $q = \{x', y'\}$ such that x' is on γ' and y' is on γ''. By Lemma 5.9, $\{q\}$ is intracyclic for $\gamma \circledast p$, and therefore, by Lemma 5.7, $\{p, q\}$ is intracyclic for γ. $\square$

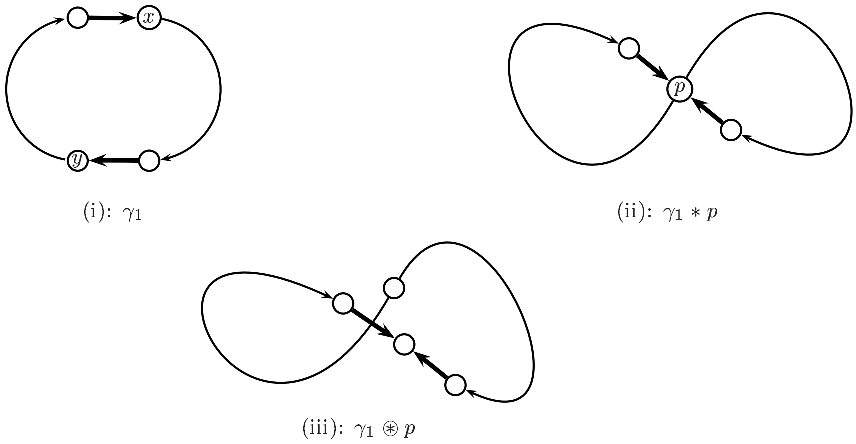


Figure 14: One pointer case for Lemma 5.9(b)

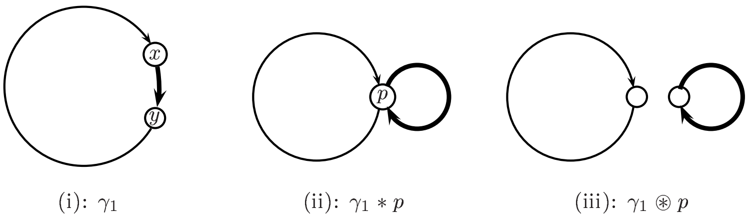


Figure 15: One pointer case for Lemma 5.9(b)

Example 5.11. Consider the genome $\mathcal{G} = (\gamma, P)$ of Fig. 17, where $P = \{p, q\}$ for the pointers $p = \{1, 6\}$ and $q = \{3, 8\}$ each of which has opposite ends. Note that $\gamma \circledast \{p, q\}$ is connected, although both $\gamma \circledast p$ and $\gamma \circledast q$ are disconnected. □

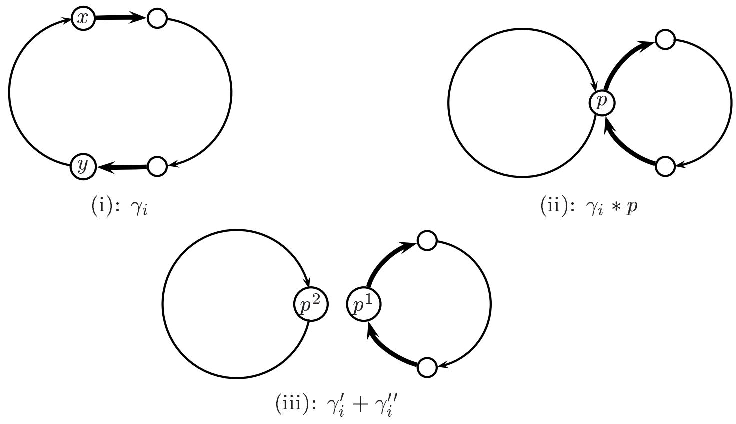
As a corollary to Lemmas 5.9 and 5.10, we have the following result.

Theorem 5.12. *For each genome $\mathcal{G}$, there exists an intracyclic assembly strategy $\mathcal{S} = (P_1, P_2, \ldots, P_m)$ such that $1 \leq |P_i| \leq 2$ for all i.*

Combining Theorem 5.12 and Theorem 5.4 one obtains for each genome that yields a connected assembled genome, an equivalent connected genome, for which there exists an intracyclic assembly strategy using no more than two pointers in each assembly step.

Theorem 5.13. *Let $\mathcal{G} = (\gamma, P)$ be a genome such that $\gamma \circledast P$ is connected. Then there exists an equivalent genome $\mathcal{G}' = (\gamma', P')$, where γ' is connected and there is an assembly strategy $\mathcal{S} = (P_1, P_2, \ldots, P_m)$ of $\mathcal{G}'$ for which $1 \leq |P_i| \leq 2$ for all i, and each $\gamma \circledast \cup_{i=1}^{j} P_i$ is a cyclic graph for $j \in [1, m]$.*

Our final example illustrates a "real life" case of gene assembly in ciliates.



(i): γ_i (ii): $\gamma_i * p$ (iii): $\gamma_i' + \gamma_i''$

Figure 16: A cyclic subgraph γ_i disconnects into γ_i' and γ_i''

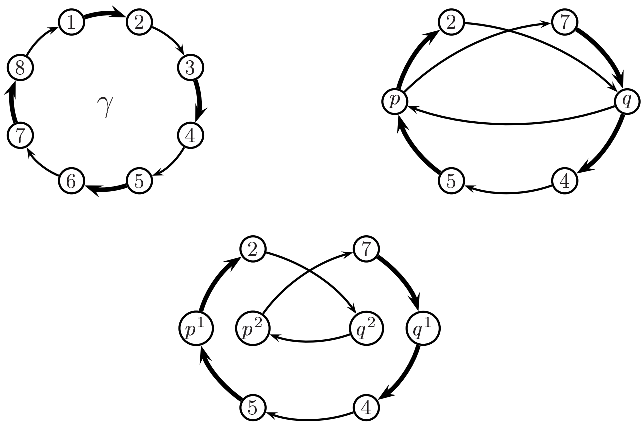


Figure 17: An intracyclic set $\{\{1,6\},\{3,8\}\}$ of pointers of Example 5.11

Example 5.14. We shall consider a gene assembly in the species *Oxytricha nova* (*O.nova*) of ciliates. The *actin I gene* in *O.nova* has nine MDSs (enumerated as $M_1, M_2, \ldots, M_9$ according to their occurrences in the macronucleus) and eight IESs (enumerated as $I_1, I_2, \ldots, I_8$ according to their occurrences in the micronucleus), see Prescott [8]. The MDS/IES structure of the micronuclear form of the actin I gene is given by the expression:

$$M_3 I_1 M_4 I_2 M_6 I_3 M_5 I_4 M_7 I_5 M_9 I_6 \overline{M}_2 I_7 M_1 I_8 M_8, \tag{6}$$

(hence the MDS M_2 is inverted). Therefore the MDSs occur in this sequence in the order $3-4-6-5-7-9-\overline{2}-1-8$. The macronuclear structure of this gene is given by the expression: $g = M_1M_2M_3M_4M_5M_6M_7M_8M_9$. In the following, each molecule M_i and I_i is considered to be a string, and the splicing of two such molecules, say $M_i \circ M_{i+1}$, is represented simply as the concatenation of these strings.

For the graphical presentation of (6) we introduce a new vertex 0 in order to make the graph cyclic, see Remark 4. In this way we obtain the cyclic graph γ of Fig. 18. Note that we have drawn the edge $(14, 13)$ with the label M_2 instead of $(13, 14)$ (with the label $\overline{M}_2$).

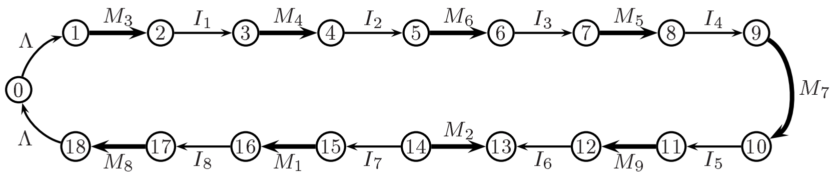


Figure 18: The cyclic graph of actin I gene in *O.nova*

The pointer set P of γ consists of

$$p_1 = \{1, 13\},\ p_2 = \{2, 3\},\ p_3 = \{4, 7\},\ p_4 = \{5, 8\},$$
$$p_5 = \{6, 9\},\ p_6 = \{10, 17\},\ p_7 = \{11, 18\},\ p_8 = \{14, 16\}.$$

The P-folded graph $\gamma * P$ is given in Fig. 19, and the final genome $\gamma \circledast P$ is given in Fig. 20, from where we can read that $M_1 M_2 \ldots M_9$ is the gene of this genome.

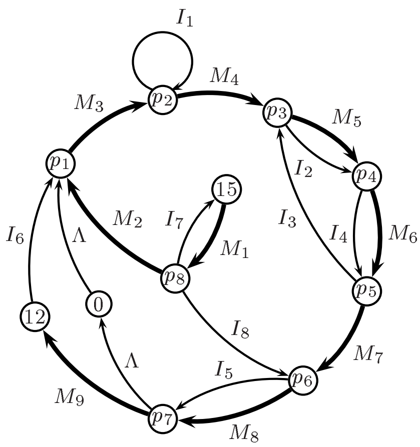


Figure 19: The P-folded graph of actin I gene in *O.nova*

Now if we remove the vertex 0 (together with its two adjacent edges), then we recover the linear structure of this component given in Fig. 21.

Hence, during the assembly of the actin I gene in *O.nova*, IES I_1 ('polluted' by a pointer) has been excised as a circular molecule, and also the combined IES $I_2 \circ I_4 \circ I_3$ (polluted by three pointers) has been excised as a circular molecule. (These circular IESs will be digested by the host cell.) The molecule containing the assembled actin I gene will also contain the combined linear IES $\overline{I}_5 \circ \overline{I}_8 \circ I_7$ (polluted by pointers) upstream from the

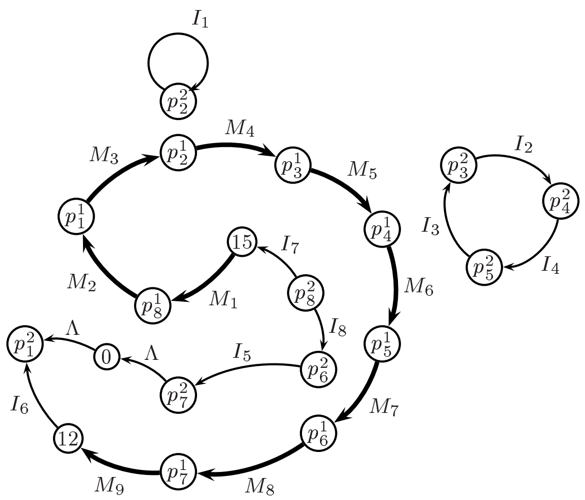


Figure 20: The final genome of actin I gene in *O.nova*

Figure 21: The component graph of actin I gene

gene, and the (polluted) IES I_6 downstream from the gene. As a matter of fact, the physical actin I gene molecule will be produced by an involved process that cuts off the molecule at the vertices 12 and 15, called *markers*, and attaches telomeres at these markers. □

Acknowledgements. Tero Harju and Grzegorz Rozenberg gratefully acknowledge support by TMR Network GETGRATS.

References

[1] Ehrenfeucht, A., Harju, T., Petre, I., Prescott, D. M., and Rozenberg, G., Formal systems for gene assembly in ciliates, *Theoret. Comput. Sci.*, to appear.

[2] Ehrenfeucht, A., Petre, I., Prescott, D. M., and Rozenberg, G., Universal and simple operations for gene assembly in ciliates, to appear in *Words, Sequences, Languages: Where computer science, biology and linguistics come across*, Mitrana, V., Martin-Vide, C. (eds.) (2000).

[3] Ehrenfeucht, A., Petre, I., Prescott, D. M., and Rozenberg, G., String and graph reduction systems for gene assembly in ciliates, *Math. Structures Comput. Sci.*, to appear.

[4] Ehrenfeucht, A., Prescott, D. M., and Rozenberg, G., Computational aspects of gene (un)scrambling in ciliates. In *Evolution as Computation*, Landwerber, L., Winfree, E. (eds.), 45–86, Springer Verlag, Berlin, Heidelberg, 2000.

[5] Kotzig, A., Moves without forbidden transitions in graph, *Mat. casopis* **18** (1968), 76 – 80.

[6] Pevzner, P. A., DNA physical mapping and alternating Eulerian cycles in colored graphs, *Algorithmica* **13** (1995), 77 – 105.

[7] Pevzner, P. A., *Computational Molecular Biology: An algorithmic approach*, MIT Press, 2000.

[8] Prescott, D. M., The evolutionary scrambling and developmental unscabling of germlike genes in hypotrichous ciliates, *Nucl. Acids Res.* **27** (1999), 1243 – 1250.

[9] Prescott, D. M., Ehrenfeucht, A., and Rozenberg, G., Molecular operations for DNA processing in hypotrichous ciliates, *European Journal of Protistology*, to appear.

[10] Tucker, A. C., A new applicable proof of the Euler circuit theorem, *Amer. Math. Monthly* **83** (1976), 638 – 640.

[11] West, D. B., *Introduction to Graph Theory*, Prentice Hall, 1996.