Sepinoud Azimi | Ion Petre

The reduction power of simple operations for gene assembly in ciliates

Turku Centre *for* Computer Science

TUCS Technical Report
No 1113, May 2014

The reduction power of simple operations for gene assembly in ciliates

Sepinoud Azimi
Åbo Akademi University, Department of Information Techology
Joukahaisenkatu 3-5 A, Turku 20520 Finland
sepinoud.azimi@abo.fi
Ion Petre
Åbo Akademi University, Department of Information Techology
Joukahaisenkatu 3-5 A, Turku 20520 Finland
ion.petre@abo.fi

TUCS Technical Report
No 1113, May 2014

Abstract

The simple intramolecular model for gene assembly in ciliates predicts correctly the assembly of all currently known ciliate gene patterns. The model consists of three molecular operations: the ld (*loop, direct-repeat excision*), the simple hi (*hairpin, inverted-repeat recombination*), and the simple dlad (*double-loop alternating direct-repeat recombination*) operations. The gene transformations conjectured by the simple intramolecular model for gene assembly can be studied as operations on the so-called directed-overlap inclusion (in short, DOI) graphs introduced in [2]. In this paper we focus on characterizing the DOI graphs that are reducible using only some combinations of the three simple operations. We also show that the DOI graph model is confluent.

Keywords: Directed overlap-inclusion graphs, gene assembly in ciliates, simple operations, confluent

TUCS Laboratory
Computational Biomodelling Laboratory

1 Introduction

Ciliates are an ancient group of unicellular eukaryotes which possess two types of nuclei: macronuclei (*MAC*) and micronuclei (*MIC*) [8, 3]. Both types of nuclei are sequences of building blocks called *macronuclear destined sequences*, or *MDS*s. In macronuclear genes the MDSs are presented in an orthodox order, whereas in micronuclear genes they are shuffled, as well as seperated by non-coding sequences called *internally eliminated sequences*, or *IES*s. During the process of sexual conjugation, the old macronucleus disintegrates and a new one is developed from the micronucleus. In this process, the micronuclear genes get transformed to their macronuclear form by having their IESs removed and their MDSs sorted in the orthodox order; this is facilitated by some short, specific nucleotide sequences called *pointers* that are repeated at the end of an MDS and at the beginning of the following MDS in macronuclear gene. This process is called *gene assembly*, see [8, 13].

We focus in this paper on the intramolecular model for gene assembly in ciliates, introduced in [9, 14]. The model consists of three molecular operations, ld, hi, and dlad, all conjecturing the folding of the gene in a specific pattern (a loop, a hairpin, or a double loop) so that a pair of pointers (two pairs in the case of dlad) are aligned. We illustrate the folding and recombinations involved in each operations in Figure 1 and refer for details on these operations to [8, 3]. The simple version of the intramolecular model, introduced in [10], assumes that the folds involved in each of the three operations are as simple as possible: in-between the aligned pointers there is a minimal number of other pointers (zero in the case of simple ld and simple dlad, one in the case of simple hi). The resulting model was shown in [10] to be capable of explaining the assembly of all currently known ciliate genes in [7]; we refer to [8] and [3] for more details. Various modeling frameworks have been proposed to represent the gene assembly in ciliates, ranging from signed permutations [11] to a string rewriting system [5, 6] to graph-based models [4, 1]. In this study, we investigate the reduction power of simple operations on the latest of such representations called *directed overlap inclusion (DOI) graphs* introduced in [1].

2 Preliminaries

We introduce here some of the notions and notations we use throughout the paper. For more details we refer to [8].

2.1 Legal strings

Let $\Delta_k = \{2, 3, \ldots, k\}$, for some $k \geq 1$, be an alphabet whose letters are called *pointers*, and let $\Sigma_k = \Delta_k \cup \{m\}$, where letter m refers to the (begining and

ending) marker. We also denote the signed copy of Σ_k by $\overline{\Sigma}_k = \overline{\Delta}_k \cup \{\overline{m}\}$, where $\Sigma_k \cap \overline{\Sigma}_k = \emptyset$. We make the convention that $\bar{\bar{p}} = p$ for all $p \in \Sigma_k \cup \{m\}$. Let $\Sigma_k^{\maltese} = (\Sigma_k \cup \overline{\Sigma}_k)^*$.

String $u \in \Sigma_k^{\maltese}$ is called a *legal string* if u contains two occurrences from the marker set $\{m, \bar{m}\}$ and for any $p \in \Delta_k$, u contains either 0 or 2 occurrences from the set $\{p, \bar{p}\}$. We define the *domain* of u as $\mathsf{dom}(u) = \{p \in \Sigma_k \mid$ either p or $\bar{p}$ occurs in $u\}$. We say that u is *sorted* if $\mathsf{dom}(u) = \{m\}$.

Let $p \in \Sigma_k \cup \bar{\Sigma}_k$ and let $u \in \Sigma_k^{\maltese}$ be a legal string, we say that p is *positive* if p and $\bar{p}$ occurs in u, and we say that it is *negative* otherwise. Let $u = u_1 p' u_2 p'' u_3$, where u_1, u_2 and u_3 are strings over $\Sigma_k \cup \overline{\Sigma}_k$ and $p', p'' \in \{p, \overline{p}\}$; substring u_2 is called the p-*interval* of u.

For any distinct $p, q \in \mathsf{dom}(u)$, p and q have one of the following relations:

- p and q *overlap* in u if exactly one occurrence from $\{p, \overline{p}\}$ can be found in the q-interval of u. We denote the overlap relation by $p \Rightarrow_u q$, if the first occurrence of p occurs in u before the first occurrence of q, and we denote it by $q \Rightarrow_u p$ otherwise;
- p *includes* q if the two occurrences from $\{q, \overline{q}\}$ are found within the p-interval. This relation is denoted by $p \rightarrow_u q$;
- p and q are *disjoint* in u if they do not overlap and neither is included in the other in u.

A gene can be represented as a legal string through its sequence of pointers and markers, for example, the legal string corresponding to actin I gene in Sterkiella nova is $34456756789m\bar{3}\,\bar{2}m289$, see [8].

The three molecular operations are formulated as rewriting rules on legal strings as follows.

Definition 1 ([5])**.** The *string pointer reduction system* is formalized as follows. In each case $p, q \in \Delta_k$ are distinct pointers and $u_1, u_2, u_3 \in \Sigma^{\maltese}$.

i. The *simple string negative* rule ssn_p is defined as follows:

$$\mathsf{ssn}_p(u_1 \tilde{p} \tilde{p} u_2) = u_1 u_2,$$

where $\tilde{p} \in \{p, \bar{p}\}$. We denote $\mathsf{Ssn} = \{\mathsf{ssn}_p \mid p \geq 2\}$.

ii. The *simple string positive* rule ssp_p is defined as follows:

$$\mathsf{ssp}_p(u_1 \tilde{p} \tilde{q} \bar{\tilde{p}} u_3) = u_1 \bar{\tilde{q}} u_3,$$

where $\tilde{p} \in \{p, \bar{p}\}$ and $\tilde{q} \in \{q, \bar{q}\}$. We denote $\mathsf{Ssp} = \{\mathsf{ssp}_p \mid p \geq 2\}$.

iii. The *simple string double* rule $\mathsf{ssd}_{p,q}$ is defined as follows:

$$\mathsf{ssd}_{p,q}(u_1 \tilde{p} \tilde{q} u_3 \tilde{p} \tilde{q} u_5) = u_1 u_3 u_5,$$

where $\tilde{p} \in \{p, \bar{p}\}$ and $\tilde{q} \in \{q, \bar{q}\}$. We denote $\mathsf{Ssd} = \{\mathsf{ssd}_{p,q} \mid p, q \geq 2\}$.

Let $\phi = \phi_r \circ \phi_{r-1} \circ \ldots \circ \phi_1$ be a composition of rules $\phi_i \in \mathsf{Ssn} \cup \mathsf{Ssp} \cup \mathsf{Ssd}$, for all $1 \leq i \leq r$; we call any such composition a *reduction strategy*. We say that ϕ is *successful* for legal string u if $\phi(u) = mm$ or $\phi(u) = \bar{m}\bar{m}$.

Example 1. Let $u = 34456756789m\bar{3}\,\bar{2}m289$ be the legal string corresponding to actin I gene in Sterkiella nova. Then $\mathsf{ssp}_3 \circ \mathsf{ssn}_5 \circ \mathsf{ssn}_4 \circ \mathsf{ssp}_2 \circ \mathsf{ssd}_{8,9} \circ \mathsf{ssd}_{6,7}$ is a reduction strategy for u:

$$
\begin{aligned}
u_1 &= \mathsf{ssd}_{6,7}(u) = 3445589m\bar{3}\,\bar{2}m289,\\
u_2 &= \mathsf{ssd}_{8,9}(u_1) = 34455m\bar{3}\,\bar{2}m2,\\
u_3 &= \mathsf{ssp}_2(u_2) = 34455m\bar{3}\,\bar{m},\\
u_4 &= \mathsf{ssn}_4(u_3) = 355m\bar{3}\,\bar{m},\\
u_5 &= \mathsf{ssn}_5(u_4) = 3m\bar{3}\,\bar{m},\\
u_6 &= \mathsf{ssp}_3(u_5) = \bar{m}\bar{m}.
\end{aligned}
$$

2.2 Overlap-inclusion graphs

Overlap-inclusion graphs (in short, OI graphs) have been introduced in [4]. Using our notation (which is slightly different than that of [4]), for a legal string u its overlap-inclusion graph $G_u = (V, \sigma, E)$ is defined as follows:

- $V = \mathsf{dom}(u)$;
- $\sigma : V \to \{+, -\}$ is the signing of vertices: for each $p \in V$, $\sigma(p) = +$ if p is a positive pointer in u and $\sigma(p) = -$ otherwise;
- $E = \{\{p, q\} \mid p \Rightarrow_u q \text{ or } q \Rightarrow_u p)\} \bigcup \{(p, q) \mid p \to_u q\}$.

In other words, for any pair of overlapping pointers $\{p, q\}$ in u there is an undirected (overlap) edge in G between p and q, and for any pointer q whose interval is included in the interval of some pointer p, G has the directed (inclusion) edge $p \to_G q$.

The simple negative rule and the simple positive rule can be defined on OI graphs so that they correspond to the string pointer reduction system; for details we refer to [4]. The simple double rule could not be defined for such graphs. The rule could however be introduced, see [1], on an extension of these graphs, the DOI graphs, that we discuss in the next session.

Example 2. Let u be the legal string corresponding to actin I gene in Sterkiella nova. Its corresponding OI graph is given in Figure 2(a).

2.3 Directed overlap-inclusion graphs

We recall here the notion of *directed overlap-inclusion (DOI) graphs* introduced in [1] and recall some of their basic properties that will be needed in this paper. We first recall that a directed graph G is called *connected* if for any distinct vertices u and v of G, there is either a (directed) path from u to v, or a (directed) path from v to u. We also recall that for a set of vertices U of G we denote by $G \setminus U$ the graph obtained from G by removing all vertices in U and all edges incident to them.

Definition 2 ([1])**.** Let u be a legal string over some Σ_k. The *directed overlap-inclusion (DOI) graph* $G_u = (V, \sigma, E_o, E_i)$ *corresponding to* u is defined as follows:

- $V = \mathsf{dom}(u)$ is the set of vertices;
- $\sigma : V \to \{+, -\}$ is the signing of vertices such that for each $p \in V$, $\sigma(p) = +$ if p is a positive pointer in u and $\sigma(p) = -$ otherwise;
- E_o and E_i are sets of its directed edges, $E_o = \{(p, q) \mid p \Rightarrow_u q\}$ and $E_i = \{(p, q) \mid p \rightarrow_u q\}$.

For a DOI graph G and any string u such that $G = G_u$ we say that G *corresponds to* u.

Two DOI graphs G and H have the same *structure*, denoted by $G \equiv H$, if there is a graph isomorphism between G and H.

Example 3. Let u be the legal string corresponding to actin I gene in Sterkiella nova; its corresponding directed overlap-inclusion graph is given in Figure 2(b).

Theorem 1 ([1])**.** *Any DOI graph G is a directed acyclic graph.*

Definition 3 ([2])**.** Let G be a DOI graph and p an arbitrary vertex of G. We introduce the following terms:

i. *Incoming inclusion edges*: we denote by $\mathsf{inSet}_i(p)$ the set of all vertices q such that $q \rightarrow p$ is an (inclusion) edge in G. Also, $\mathsf{inDeg}_i(p)$ is the number of vertices in $\mathsf{inSet}_i(p)$.

ii. *Outgoing inclusion edges*: we denote by $\mathsf{outSet}_i(p)$ the set of all vertices q such that $p \rightarrow q$ is an (inclusion) edge in G. Also, $\mathsf{outDeg}_i(p)$ is the number of vertices in $\mathsf{outSet}_i(p)$.

iii. *Incoming overlap edges*: we denote by $\mathsf{inSet}_o(p)$ the set of all vertices q such that $q \Rightarrow p$ is an (overlap) edge in G. Moreover, $\mathsf{inDeg}_o(p)$ is the number of vertices in $\mathsf{inSet}_o(p)$.

iv. *Outgoing overlap edges*: we denote by $\mathsf{outSet}_o(p)$ the set of all vertices q such that $p \Rightarrow q$ is an (overlap) edge in G. Also, $\mathsf{outDeg}_o(p)$ is the number of vertices in $\mathsf{outSet}_o(p)$.

We recall now the definition of simple operations for DOI graphs introduced in [2]. The operations are defined in general for any directed, vertex- and edge-labeled graph, in particular for DOI graphs.

Definition 4 ([2])**.** Let $G = (V, \sigma, E_o, E_i)$ be a directed, vertex- and edge-labeled graph. For any distinct vertices $p, q \in V \setminus \{m\}$, the graph operations sgn_p, sgp_p and $\mathsf{sgd}_{p,q}$ are defined on G as follows:

(i) *The simple graph negative rule* sgn for p, denoted sgn_p, is applicable to G if:

- $\sigma(p) = -$ and
- $\mathsf{inDeg}_o(p) = \mathsf{outDeg}_o(p) = \mathsf{outDeg}_i(p) = 0$.

In this case, $\mathsf{sgn}_p(G) = G \setminus \{p\}$. We denote $\mathsf{Sgn} = \{\mathsf{sgn}_p \mid p \in \Delta_k, p \geq 2\}$. We say that sgn_p *corresponds* to the string-rewriting rule ssn_p.

(ii) *The simple graph positive rule* sgp for p, denoted sgp_p, is applicable to G if:

- $\sigma(p) = +$,
- $\mathsf{inDeg}_o(p) + \mathsf{outDeg}_o(p) = 1$, and
- $\mathsf{outDeg}_i(p) = 0$.

Let q be the vertex with the property $\mathsf{inSet}_o(p) \cup \mathsf{outSet}_o(p) = \{q\}$. In this case, $\mathsf{sgp}_p(G)$ is the graph obtained from $G \setminus \{p\}$ by switching the label of q: q is negative in $\mathsf{sgp}_p(G)$ if and only if it is positive in G. We denote $\mathsf{Sgp} = \{\mathsf{sgp}_p \mid p \in \Delta_k, p \geq 2\}$. We say that sgp_p *corresponds* to the string-rewriting rule ssp_p.

(iii) *The simple graph double rule* sgd for p, q, denoted $\mathsf{sgd}_{p,q}$, is applicable to G if:

- $\sigma(p) = \sigma(q) = -$,
- $q \in \mathsf{outSet}_o(p)$,
- $\mathsf{inSet}_o(p) \cup \{p\} = \mathsf{inSet}_o(q)$,
- $\mathsf{outSet}_o(p) = \mathsf{outSet}_o(q) \cup \{q\}$,
- $\mathsf{inSet}_i(p) = \mathsf{inSet}_i(q)$ and
- $\mathsf{outSet}_i(p) = \mathsf{outSet}_i(q)$.

In this case, $\mathsf{sgd}_{p,q}(G) = G \setminus \{p, q\}$. We denote $\mathsf{Sgd} = \{\mathsf{sgd}_{p,q} \mid p, q \in \Delta_k, p, q \geq 2, p \neq q\}$. We say that $\mathsf{sgd}_{p,q}$ corresponds to the string-rewriting rule $\mathsf{ssd}_{p,q}$.

A *reduction strategy* is a composition of simple graph rules $\phi = \phi_{p_n} \circ \ldots \circ \phi_{p_2} \circ \phi_{p_1}$. We say that ϕ is *successful* on G if $\phi(G)$ is the graph having only vertex m where m is negative. Let $\Omega \subseteq \{\mathsf{Sgn}, \mathsf{Sgp}, \mathsf{Sgd}\}$ be a set of types of graph rules. If all rules in ϕ are from the rule sets indicated by Ω, then we say that ϕ is an Ω-*strategy* and that ϕ is Ω-*successful*, resp. We also say that a directed, vertex- and edge-labeled graph is Ω-*reducible* if it has an Ω-successful reduction strategy.

Example 4. Let G be the DOI graph corresponding to actin I gene in Sterkiella nova. A successful reduction strategy of G is presented in Figure 3.

We recall the equivalence between the string-based model for simple gene assembly and the DOI graph-based model.

Theorem 2 ([2])**.** *Let u be a legal string, G_u its corresponding DOI graph. Let also $\phi \in \mathsf{Ssn} \cup \mathsf{Ssp} \cup \mathsf{Ssd}$ and $\psi \in \mathsf{Sgn} \cup \mathsf{Sgp} \cup \mathsf{Sgd}$ be the DOI graph rule corresponding to ϕ. Then ϕ is applicable to u if and only if ψ is applicable to G_u. In this case, $G_{\phi(u)} = \psi(G_u)$.*

The following result is straightforward from the definition of our operations and that of the DOI graph equivalence.

Lemma 3. *Let G, H be two DOI graphs such that $G \equiv H$; let $\mu : G \to H$ be the graph isomorphism between them.*

i. If sgn_p is applicable to G, then $\mathsf{sgn}_{\mu(p)}$ is applicable to H and $\mathsf{sgn}_p(G) \equiv \mathsf{sgn}_{\mu(p)}(H)$.

ii. If sgp_p is applicable to G, then $\mathsf{sgp}_{\mu(p)}$ is applicable to H and $\mathsf{sgp}_p(G) \equiv \mathsf{sgp}_{\mu(p)}(H)$.

iii. If $\mathsf{sgd}_{p,q}$ is applicable to G, then $\mathsf{sgd}_{\mu(p),\mu(q)}$ is applicable to H and $\mathsf{sgd}_{p,q}(G) \equiv \mathsf{sgd}_{\mu(p),\mu(q)}(H)$.

3 The reduction power of the simple operations

In this section we focus on characterizing the DOI graphs that are reducible using simple operations of different types.

3.1 $\{\mathsf{Sgn}\}$-reducible graphs

Theorem 4. *Let $G = (V, \sigma, E_o, E_i)$ be a DOI graph. G is Sgn-reducible if and only if $E_o = \emptyset$ and $\sigma(v) = -$, for all $v \in V$.*

Proof. We first prove the reverse implication. Let G be a DOI graph with a single negative node; clearly G is reducible through Sgn operations. Let $n \geq 2$ and assume that all DOI graphs of size $n-1$ with only inclusion edges and negative vertices are reducible using only sgn. Let G be such a DOI graph of size n. By Theorem 1, G is acyclic and so there is a node p such that $\mathsf{inDeg}_i(p) = 0$. Thus, $\mathsf{sgn}_p(G)$ is applicable to G; let $G' = \mathsf{sgn}_p(G) = G \setminus \{p\}$. The conclusion follows by applying the induction hypothesis to G'.

The proof of the direct implication is straightforward. Let G be a DOI graph reducible using only Sgn. By Definition 4, applying an Sgn operation neither removes any overlap edge, nor changes any signing of the vertices. Therefore all edges are inclusion edges and all vertices are negative. □

3.2 $\{\mathsf{Sgn}, \mathsf{Sgp}\}$-reducible graphs

The next theorem characterizes the reduction power of $\mathsf{Sgn} \cup \mathsf{Sgp}$-operations. The formulation and the proof of the result follows closely the corresponding result of [4] for OI graphs. We first introduce some notations.

For a composition $\phi = \phi_{p_n} \circ \ldots \circ \phi_{p_2} \circ \phi_{p_1}$ of $\mathsf{Sgn} \cup \mathsf{Sgp}$-operations ϕ_{p_i}, $1 \leq i \leq n$, we denote by $\mathsf{dom}_-(\phi) = \{p \in \Delta_k \mid \exists 1 \leq i \leq n \text{ such that } \phi_{p_i} = \mathsf{sgn}_p\}$. We also denote $\mathsf{ord}(\phi) = (p_1, p_2, \ldots, p_n)$ the order in which the vertices are reduced by ϕ.

For a directed graph G we say that an ordering $P = (p_1, p_2, \ldots, p_n)$ of its vertices is *anti-topological* if there is no edge from p_i to p_j for all $i < j$. In particular, in the case of a DOI graph, an anti-topological ordering of its vertices takes into account both its inclusion and its overlap edges.

For a DOI graph $G = (V, \sigma, E_o, E_i)$, we define $G_o = (V, \sigma, E'_o)$ to be the *undirected* subgraph of G induced by its overlap edges: $E'_o = \{\{i, j\} \mid (i, j) \in E_o \text{ or } (j, i) \in E_o\}$. For a vertex p, $\deg_{G_o}(p)$ denotes its degree in the (undirected) graph G_o.

Theorem 5. *Let $G = (V, \sigma_G, E_o, E_i)$ be a DOI graph. Let $N \subseteq V \setminus \{m\}$ and $P = (p_1, p_2, \ldots, p_n, p_{n+1})$ a linear ordering of V with $p_{n+1} = m$. There is an $\{\mathsf{Sgn}, \mathsf{Sgp}\}$-successful reduction ϕ of G with $N = \mathsf{dom}_-(\phi)$ and $P = (\mathsf{ord}(\phi), m)$ if and only if the following conditions are satisfied:*

i. G_o is a forest;

ii. For each vertex $q \in G$, $\deg_{G_o}(q)$ is even if and only if $\sigma_G(q) = -$;

iii. Each tree in the forest G_o has exactly one vertex in $N \cup \{m\}$;

iv. Consider G_o as a rooted forest with its roots in $N \cup \{m\}$ and denote by G_N the graph obtained from G by changing the orientation of all its edges $s \Rightarrow t$ where s is a child of t in the rooted forest G_o. Then G_N is acyclic and P is an anti-topological ordering of G_N.

Proof. We prove the result by showing the equivalence of both sides of the statement with the following set of conditions:

a. For all $1 \leq i < j \leq n+1$, there is no inclusion edge $p_i \to p_j$ in G;

b. For all $1 \leq i \leq n$, if $p_i \in N$, then there is no overlap edge between p_i and p_j in G, in either direction, for any $1 \leq i < j \leq n$. If $p_i \notin N$, then there is exactly one $j > i$ such that there exists an overlap edge between p_i and p_j in G.

c. For each vertex $q \in G$, $\deg_{G_o}(q)$ is even if and only if $\sigma_G(q) = -$.

Claim 1. *The left-hand side of the theorem's statement holds if and only if conditions a-c are satisfied.*

Proof of Claim 1: Let $\phi = \phi_{p_n} \circ \ldots \circ \phi_{p_1}$ be an $\{\mathsf{Sgn}, \mathsf{Sgp}\}$-successful strategy, $N = \mathsf{dom}_-(\phi)$ and $P = (\mathsf{ord}(\phi), m) = (p_1, p_2, \ldots, p_n, p_{n+1})$, with $p_{n+1} = m$. Denote $G_i = \phi_{p_i} \circ \ldots \circ \phi_{p_1}(G)$, for all $1 \leq i \leq n$.

Let $1 \leq i < j \leq n+1$, i.e., p_i is reduced before p_j either by an sgn or an sgp operation. By Definition 4 we can conclude that there is no outgoing inclusion edge from p_i to p_j in G_{i-1}. Since G_{i-1} is obtained from G by successive node removals and possible node label switches, it follows that the same is true in G, thus proving condition **a**.

Let $1 \leq i \leq n$. If $p_i \in N = \mathsf{dom}_-(\phi)$, then by the definition of sgn operation it is easy to see that p_i has no overlap edge incident to any vertex p_j with $j > i$. If $p_i \notin N$, then p_i is reduced in ϕ through sgp_{p_i}. Therefore, by Definition 4, there is only one vertex p_j adjacent to p_i in G_{i-1} with an overlap edge where $i < j$. The same clearly holds also in G, thus proving condition **b**.

Condition **c** can easily be obtained through induction on the number of vertices of G.

To prove the reverse implication we use induction on the number of vertices of graph G.

If $n = 0$, then $p_1 = m$ and $\deg_{G_o}(p_1) = 0$. By (c) we have $\sigma(p_1) = -$ and so, G is trivially reducible through the empty reduction strategy. Assume now that the claim holds for all graphs with at most k vertices, for some $k \geq 1$. Let G be a DOI graph with $k+1$ vertices, that satisfies conditions **a-c**.

If $p_1 \in N$, then by **a-b** it is either isolated or it has some incoming inclusion edges; therefore $\deg_{G_o}(p_1) = 0$, $\sigma(p_1) = -$ and so, sgn_{p_1} is applicable to G. Consequently, based on the definition of sgn, $G_1 = \mathsf{sgn}_{p_1}(G)$ is a DOI graph with k vertices that satisfies conditions **a-c**. By the induction hypothesis G_1 is $\{\mathsf{Sgn}, \mathsf{Sgp}\}$-reducible and thus, so is G.

If $p_1 \notin N$, by condition **a** p_1 has no outgoing inclusion edges and by **b**, there is exactly one $j > 1$ such that there is an overlap edge between p_1 and p_j, and by **c**, $\sigma(p_1) = +$. Hence, p_1 is reducible using sgp. Applying sgp does not add any edges and it only changes the signing of vertex p_j. Consequently, $G_1 =$

$\mathsf{sgp}_{p_1}(G)$ is a DOI graph with k vertices that satisfies conditions **a-c**. By induction hypothesis, G_1 is $\{\mathsf{Sgn}, \mathsf{Sgp}\}$-reducible and thus, so is G.

Claim 2. *The right-hand side of the theorem's statement holds if and only if conditions **a-c** are satisfied.*

Proof of Claim 2: We note immediately that condition **c** is identical to ii.

We first prove the direct implication. By iv P is an anti-topological ordering of G_N. Therefore for every $i < j$, there is no edge from p_i to p_j in G_N. Since the inclusion edges in G_N are the inclusion edges in G, it follows that for all $i < j$ there is no inclusion edge $p_i \rightarrow p_j$ in G, proving **a**.

Let $1 \leq i \leq n$. If $p_i \in N$, then p_i is the root of a rooted tree in the forest G_o. By iv, it follows that if there is any overlap edge incident to p_i in G_N, then it is an outgoing overlap edge from p_i. Therefore, for every $j > i$ if there is an overlap edge between p_i and p_j in G, then G_N has an edge from p_i to p_j, which contradicts P being an anti-topological ordering of G_N. Hence, for $p_i \in N$, there is no overlap edge between p_i and p_j for any $j > i$. This proves the first part of **b**.

If $p_i \notin N$, then p_i is not the root of a rooted tree in G_o and so, there exists a directed edge $p_j \Rightarrow p_i$ in G_N. Since P is an anti-topological ordering on the vertices of G_N, it follows that $i < j$. Assume now that there are $j_1 > j_2 > i$ such that there is an overlap edge between p_{j_1} to p_i and an overlap edge between p_{j_2} and p_i in G. This implies that $p_{j_1} \Rightarrow_{G_N} p_i$ and $p_{j_2} \Rightarrow_{G_N} p_i$ are edges of G_N, based on the anti-topological ordering of its vertices. Observe now that p_{j_1} and p_{j_2} are vertices in the same rooted tree of G_N and so, there is a directed path ρ_1 from its root, say p_r, to p_{j_1} and a directed path ρ_2 from p_r to p_{j_2}. Note however that ρ_1, ρ_2 and the edges $p_{j_1} \Rightarrow_{G_N} p_i$, $p_{j_2} \Rightarrow_{G_N} p_i$ induce a cycle in G_o, contradicting i. This concludes the proof of the second part of **b**.

To prove the reverse implication, assume that G_o has a cycle and let p_i be the vertex with the smallest index on that cycle. This implies that there are two vertices $j_1, j_2 > i$ such that there is an edge between p_i and p_{j_1} and an edge between p_i and p_{j_2}. This contradicts **b**, thus proving i.

To prove iii, let T be a tree in G_o and let p_i be the vertex in T with the largest index i. By **b**, if $p_i \notin N \cup \{m\}$, there is an overlap edge between p_i and a vertex p_j where $j > i$, a contradiction with our choice of i.

Assume now there are two vertices of T $p_i, p_j \in N \cup \{m\}$, where $i < j$. There is an overlap path, say ρ, in T from p_i to p_j: $\rho = (p_{k_1}, \ldots, p_{k_r})$, where $r \geq 2$, $k_1 = i$ and $k_r = j$. By **b**, p_i has no overlap edge incident to any vertex with a larger index and so, $k_2 < k_1$. It also follows by **b** that p_{k_2} can have at most one overlap edge incident to a vertex with a larger index, which is p_{k_1}; therefore, $k_3 < k_2$. Iterating this argument we conclude that $j = k_r < \ldots < k_2 < k_1 = i$, contradicting that $i < j$.

To prove iv, let p_i, p_j be two vertices in G where $i < j$ such that there is an overlap edge $p_i \Rightarrow_{G_N} p_j$. Take i to be the smallest such index. Then by **b**, $p_i \notin N$. By the construction of G_N, it follows that there is another vertex p_k such that

$p_k \Rightarrow_{G_N} p_i$. If $k > i$, then p_i is overlap-adjacent to two vertices with index larger than i, which contradicts **b**. Thus, $k < i$; this however contradicts our choice of i. Hence P is an anti-topological ordering of G_N. It then follows that G_N is also acyclic. □

3.3 {Sgp}-reducible graphs

Theorem 6. *Let $G = (V, \sigma_G, E_o, E_i)$ be a DOI graph with $m \in V$. G is {Sgp}-reducible if and only if the following conditions are satisfied:*

i. G_o is a tree;

ii. For each vertex $q \in G$, $\deg_{G_o}(q)$ is even if and only if $\sigma_G(q) = -$;

iii. Let G_N be the graph obtained from G by changing the orientation of all its directed edges $s \Rightarrow t$ where s is a child of t in the rooted forest G_o. Then G_N is acyclic and each successful reduction in G corresponds to an anti-topological ordering of G_N.

Proof. Note that Theorem 5 implies an ordering in which the graph is reduced and the vertices in N are the last ones to be reduced in every connected component of the graph. Therefore $N = \emptyset$ in the case where a DOI graph is reduced using only Sgp. Then the claim follows by Theorem 5. □

Example 5. Let G be the DOI graph corresponding to actin I gene in Sterkiella nova given in Figure 2(b). Since G_o is not a forest, it follows by Theorem 5 that G is not {Sgn, Sgp}-reducible. Let G' be the induced subgraph of G given in Figure 4(a); let $N = \{4, 5\}$ and $P = (2, 5, 4, 3)$. It follows by Theorem 5 that G' is {Sgn, Sgp}-reducible with the successful strategy $\phi = \mathsf{sgp}_3\,\mathsf{sgn}_4\,\mathsf{sgn}_5\,\mathsf{sgp}_2(G')$.

Note that G' is not {Sgp}-reducible, since G'_o is not a tree and therefore, G' does not satisfy the conditions in Theorem 6. On the other hand, but the subgraph G'' given in Figure 4(b) is {Sgp}-reducible.

The following problems concerning the reduction power of the simple operations remain open:

i. characterize {Sgd}-reducible DOI graphs;

ii. characterize {Sgn, Sgd}-reducible DOI graphs;

iii. characterize {Sgp, Sgd}-reducible DOI graphs;

iv. characterize {Sgn, Sgp, Sgd}-reducible DOI graphs.

4 Confluent strategies on DOI graphs

It has been shown in [12] and [5] that the strategies using simple operations are confluent for signed permutations and legal strings. In this section we show that this result holds also for DOI graphs. The results here are similar to those of [12] and [5] in the case of permutations and strings.

Definition 5. Let $G = (V, \sigma_G, E_o, E_i)$ be a DOI graph. We say that the reduction strategy ϕ for G is *maximal* if either ϕ is successful for G, or no operation is applicable to $\phi(G)$. We say that two reduction strategies ϕ, ψ for G are *confluent* if $\phi(G) \equiv \psi(G)$.

Lemma 7. *Let G be a DOI graph and $\phi \in \mathsf{Sgn}$, $\psi \in \mathsf{Sgn} \cup \mathsf{Sgp} \cup \mathsf{Sgd}$ be two operations applicable to G. Then both $\phi \circ \psi$ and $\psi \circ \phi$ to G are applicable and $\phi \circ \psi(G) = \psi \circ \phi(G)$.*

Proof. The result follows easily since applying sgn on one vertex neither has any effect on the signing of the other vertices nor adds any edges to influence the applicability of any other operations. □

Lemma 8. *Let G be a DOI graph and $\phi, \psi \in \mathsf{Sgp}$ be two distinct operations applicable to G. Then either $\phi \circ \psi(G) = \psi \circ \phi(G)$ or $\phi(G) \equiv \psi(G)$.*

Proof. Let $\phi = \mathsf{sgp}_p$ and $\psi = \mathsf{sgp}_q$, $p \neq q$. We consider the following two cases:

Case 1 If vertices p and q do not overlap, then clearly, by definition, $\phi \circ \psi(G) = \psi \circ \phi(G)$.

Case 2 If vertices p and q overlap, say $p \Rightarrow q$, then the only other edges incident to p and q are incoming inclusion edges. It follows then that $\mathsf{sgn}_q \circ \mathsf{sgp}_p(G) = \mathsf{sgn}_p \circ \mathsf{sgp}_q(G)$, i.e., $\phi(G) \equiv \psi(G)$.

□

Lemma 9. *Let G be a DOI graph and $\phi, \psi \in \mathsf{Sgd}$ be two distinct operations applicable to G. Then either $\phi \circ \psi(G) = \psi \circ \phi(G)$ or $\phi(G) \equiv \psi(G)$.*

Proof. Let $\phi = \mathsf{sgd}_{p,q}$ and $\psi = \mathsf{sgd}_{r,s}$. If $\{p, q\} \neq \{r, s\}$, clearly $\phi \circ \psi(G) = \psi \circ \phi(G)$. Consider now the case where $|\{p, q\} \cap \{r, s\}| = 1$. Note that if $p = r$ or $q = s$, then $\mathsf{sgd}_{p,q}$ and $\mathsf{sgd}_{r,s}$ are not applicable simultaneously to G. The followng two cases are then possible:

Case 1 $q = r$, $p \Rightarrow q$ and $q \Rightarrow s$,

Case 2 $p = s$, $p \Rightarrow q$ and $r \Rightarrow p$.

It is enough to discuss here only Case 1, as the other is symmetric. Denote $G' = \mathsf{sgd}_{p,q}(G)$ and $G'' = \mathsf{sgd}_{q,s}(G)$. Then $\mathsf{outSet}_o(p) \setminus \{q\} = \mathsf{outSet}_o(q)$ and $\mathsf{outSet}_o(q) \setminus \{s\} = \mathsf{outSet}_o(s)$. As a result $\mathsf{outSet}_o(s)$ in G' is equal to $\mathsf{outSet}_o(p)$ in G''. Similar results hold true for the sets of outgoing inclusion edges and incoming overlap and inclusion edges. Hence, $\phi(G) \equiv \psi(G)$. □

Lemma 10. *Let G be a DOI graph and $\phi \in \mathsf{Sgp}$, $\psi \in \mathsf{Sgd}$ be two operations applicable to G. Then $\phi \circ \psi(G) = \psi \circ \phi(G)$.*

Proof. Let $\phi = \mathsf{sgp}_p$ and $\psi = \mathsf{sgp}_{q,r}$. If $p \notin \{q, r\}$, then clearly, by definition, $\phi \circ \psi(G) = \psi \circ \phi(G)$.

Let $p \in \{q, r\}$. Since sgp_p is applicable to G, p is a positive vertex in G. But $\mathsf{sgd}_{q,r}$ is also applicable to G and so, both q and r should be negative in G, a contradiction. □

Theorem 11. *Let G be a DOI graph and ϕ, ψ be two maximal strategies for G. Then ϕ and ψ are confluent.*

Proof. We prove by induction on the number of vertices of G.

The case when G has only one vertex is trivial. Suppose now the claim holds for all DOI graphs with $|V| \leq k$.

Let G be a DOI graph with $|V| = k + 1$ and $\phi = \phi_m \circ \ldots \circ \phi_2 \circ \phi_1$, $\psi = \psi_n \circ \ldots \circ \psi_2 \circ \psi_1$. If $\phi_1 = \psi_1$, the the claim follows by the induction hypothesis. Assume that they are distinct, but both applicable to G. By Lemmas 7-10, either $\phi_1 \circ \psi_1(G) = \psi_1 \circ \phi_1(G)$ or $\phi_1(G) \equiv \psi_1(G)$. The latter case is concluded based on Lemma 3.

If $\phi_1 \circ \psi_1(G) = \psi_1 \circ \phi_1(G)$, then by induction hypothesis all maximal strategies on $\phi_1(G)$ are confluent; consequently, all maximal strategies for $\psi_1 \circ \phi_1(G)$ are also confluent. Similarly, all strategies on $\psi_1(G)$ are confluent; consequently, all maximal strategies on $\phi_1 \circ \psi_1(G)$ are also confluent. Since $\phi_1 \circ \psi_1(G) = \psi_1 \circ \phi_1(G)$, it follows that all maximal strategies on $\phi_1(G)$ are confluent with all maximal strategies on $\psi_1(G)$, concluding the proof. □

Corollary 12. *Let G be a DOI graph. Then either all its maximal strategies are successful, or they are all unsuccessful and in this case, the resulting graphs have the same structure.*

References

[1] S. Azimi, T. Harju, M. Langille, I. Petre, and V. Rogojin. Directed overlap-inclusion graphs as representations of ciliate genes. *Fundamenta Informaticae*, 110(1):29–44, 2011.

[2] Sepinoud Azimi, Tero Harju, Miika Langille, and Ion Petre. Simple gene assembly as a rewriting of directed overlap-inclusion graphs. *Theoretical Computer Science*, 454:30–37, 2012.

[3] R. Brijder, T. Harju, N. Jonoska, I. Petre, and G. Rozenberg. Gene assembly in ciliates. In G. Rozenberg, T.H.W. Bäck, and J.N. Kok, editors, *Handbook of Natural Computing*, volume 2 (Molecular Computation). Springer, 2012.

[4] R. Brijder and H.J. Hoogeboom. Combining overlap and containment for gene assembly in ciliates. *Theoretical Computer Science*, 411(6):897–905, 2010.

[5] R. Brijder, M. Langille, and I. Petre. A string-based model for simple gene assembly. In E. Csuhaj-Varju and Z. Esik, editors, *Proceedings of FCT 2007*, volume 4639 of *Lecture Notes in Computer Science*, pages 161–172. Springer, 2007.

[6] R. Brijder, M. Langille, and I. Petre. Extended strings and graphs for simple gene assembly. *Theoretical Computer Science*, 411(4-5):730–738, 2010.

[7] A.R.O. Cavalcanti, T.H. Clarke, and L.F. Landweber. Mds_ies_db: a database of macronuclear and micronuclear genes in spirotrichous ciliates. *Nucleic acids research*, 33(1):D396, 2005.

[8] A. Ehrenfeucht, T. Harju, I. Petre, D.M. Prescott, and G. Rozenberg. *Computation in living cells: gene assembly in ciliates*. Springer, 2004.

[9] A. Ehrenfeucht, D.M. Prescott, and G. Rozenberg. Computational aspects of gene (un) scrambling in ciliates. In L.F. Landweber and E. Winfree, editors, *Evolution as computation*, pages 216–256. Springer, 2001.

[10] T. Harju, I. Petre, V. Rogojin, and G. Rozenberg. Patterns of simple gene assembly in ciliates. *Discrete Applied Mathematics*, 156(14):2581–2597, 2008.

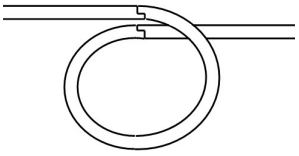
[11] T. Harju and G. Rozenberg. Computational processes in living cells: gene assembly in ciliates. *Lecture Notes in Computer Science*, 2450:1–20, 2003.

[12] M. Langille and I. Petre. Simple gene assembly is deterministic. *Fundamenta Informaticae*, 73(1):179–190, 2006.

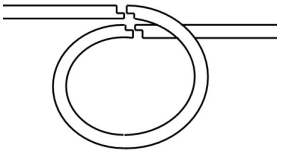
[13] I. Petre and G. Rozenberg. Gene assembly in ciliates. *Scholarpedia*, 5(1):9269, 2010.

[14] D.M. Prescott, A. Ehrenfeucht, and G. Rozenberg. Molecular operations for dna processing in hypotrichous ciliates. *European Journal of Protistology*, 37(3):241–260, 2001.

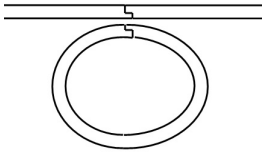
ld



(a) The molecule folds as a loop to align the pointers.

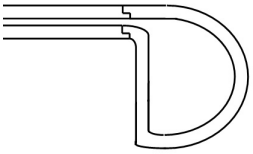


(b) Enzymes cut on the pointer sites.

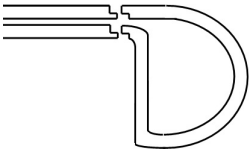


(c) As a result, one circular molecule gets excised from the original DNA molecule.

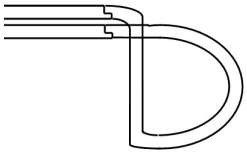
hi



(d) The molecule folds as a hairpin to align the pointers.

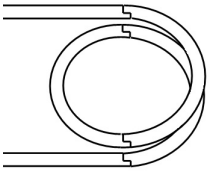


(e) Enzymes cut on the pointer sites.

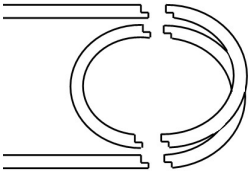


(f) As a result, a segment of the molecule gets inverted.

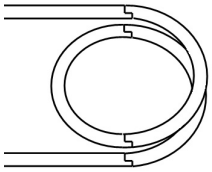
dlad



(g) The molecule folds as a double loop to align the pointers.



(h) Enzymes cut on the pointer sites.



(i) As a result, two segments of the molecule get inter-changed.

Figure 1: The three operations of the intramolecular model for gene assembly in ciliates.

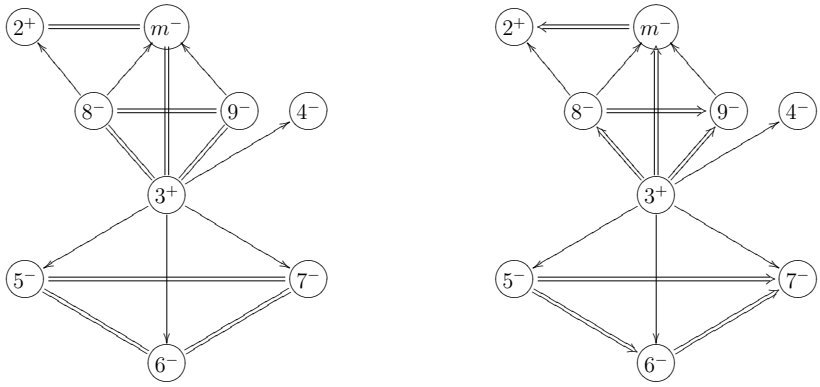


Figure 2: (a) The OI graph corresponding to actin I gene in Sterkiella nova, (b) the DOI graph corresponding to actin I gene in Sterkiella nova.

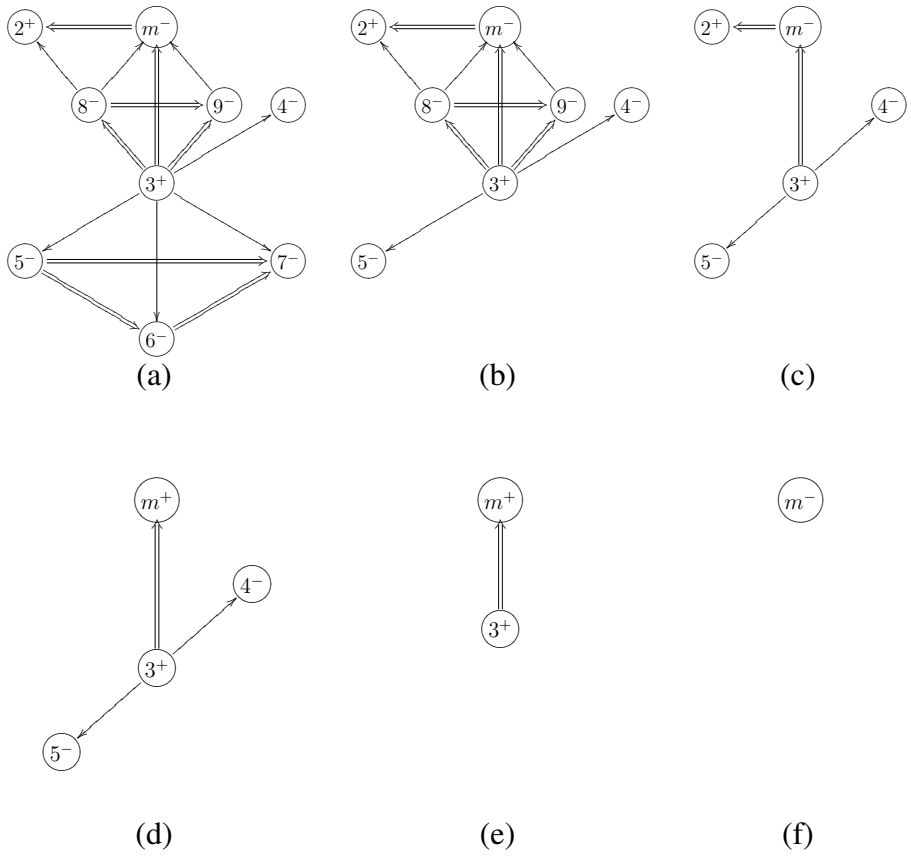
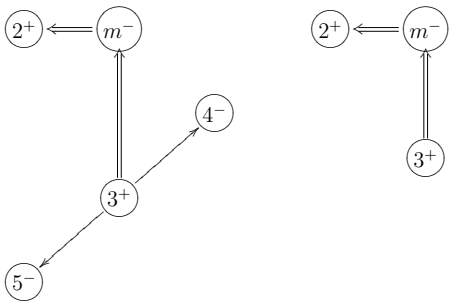


Figure 3: (a) The DOI graph G corresponding to actin I gene in Sterkiella nova; (b) $G' = \mathsf{sgd}_{6,7}(G)$; (c) $G'' = \mathsf{sgd}_{8,9}(G')$; (d) $G''' = \mathsf{sgp}_2(G'')$; (e) $G^4 = \mathsf{sgn}_4\,\mathsf{sgn}_5(G''')$; (f) $G^5 = \mathsf{sgp}_3(G^4)$.



(a) (b)

Figure 4: (a) An $\{\mathsf{Sgn}, \mathsf{Sgp}\}$-reducible subgraph of the DOI graph corresponding to actin I gene in Sterkiella nova. (b) An $\{\mathsf{Sgp}\}$-reducible subgraph of the DOI graph corresponding to actin I gene in Sterkiella nova.

Joukahaisenkatu 3-5 B, 20520 Turku, Finland | www.tucs.fi

University of Turku
- Department of Information Technology
- Department of Mathematics

Åbo Akademi University
- Department of Information Technologies

Turku School of Economics
- Institute of Information Systems Sciences

ISBN 978-952-12-3081-3
ISSN 1239-1891