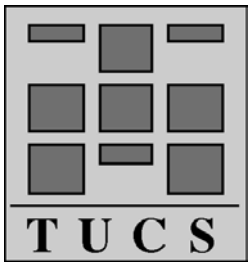
Two models for gene assembly in ciliates

Tero Harju
Ion Petre
Grzegorz Rozenberg



Turku Centre for Computer Science

TUCS Technical Reports

No 604, April 2004

Two models for gene assembly in ciliates

Tero Harju

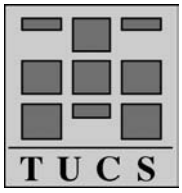
Department of Mathematics, University of Turku and
Turku Centre for Computer Science
Turku 20014 Finland
harju@utu.fi

Ion Petre

Department of Computer Science, Åbo Akademi University and
Turku Centre for Computer Science
Turku 20520 Finland
ipetre@abo.fi

Grzegorz Rozenberg

Leiden Institute for Advanced Computer Science, Leiden University
Niels Bohrweg 1, 2333 CA Leiden, the Netherlands, and
Department of Computer Science, University of Colorado at Boulder
Boulder, Co 80309-0347, USA
rozenber@liacs.nl



Turku Centre for Computer Science
TUCS Technical Report No 604
April 2004
ISBN 952-12-1335-3
ISSN 1239-1891

Abstract

Two models for gene assembly in ciliates have been proposed and investigated in the last few years. The DNA manipulations postulated in the two models are very different: one model is *intramolecular* – a single DNA molecule is involved here, folding on itself according to various patterns, while the other is *intermolecular* – two DNA molecules may be involved here, hybridizing with each other. Consequently, the assembly strategies predicted by the two models are completely different. Interestingly however, the final result of the assembly (including the assembled gene) is always the same. We compare in this paper the two models for gene assembly, formalizing both in terms of pointer reductions. We also discuss invariants and universality results for both models.

TUCS Laboratory
Discrete Mathematics for Information Technology Laboratory

1 Introduction

Ciliates are unicellular eukaryotic organisms, see, e.g. [23]. This is an ancient group of organisms, estimated to have originated around two billion years ago. It is also a very diverse group – some 8000 species are currently known and many others are likely to exist. Their diversity can be appreciated by comparing their genomic sequences: some ciliate types differ genetically more than humans differ from fruit flies! Two characteristics unify ciliates as a single group: the possession of hairlike cilia used for motility and food capture, and the presence of two kinds of functionally different nuclei in the same cell, a micronucleus and a macronucleus, see [15], [24], [25]; the latter feature is unique to ciliates. The macronucleus is the "household" nucleus – all RNA transcripts are produced in the macronucleus. The micronucleus is a germline nucleus and has no known function in the growth or in the division of the cell. The micronucleus is activated only in the process of sexual reproduction, where at some stage the micronuclear genome gets transformed into the macronuclear genome, while the old macronuclear genome is destroyed. This process is called *gene assembly*, it is the most involved DNA processing known in living organisms, and it is most spectacular in the *Stichotrichs* species of ciliates (which we consider in this paper). What makes this process so complex is the unusual rearrangements that ciliates have engineered in the structure of their micronuclear genome. While genes in the macronucleus are contiguous sequences of DNA placed (mostly) on their own molecules (and some of them are the shortest DNA molecules known in Nature), the genes in the micronucleus are placed on long chromosomes and they are broken into pieces called *MDSs*, separated by noncoding blocks called *IESs*, see [15, 21, 22, 23, 24, 25, 26]. Adding to the complexity, the order of the MDSs is permuted and MDSs may be inverted. One of the amazing features of this process is that ciliates appear to use "linked lists" in gene assembly, see [29, 30], similarly as in software engineering!

Two different models have been proposed for gene assembly. The first one, proposed by Landweber and Kari, see [19], [20], is intermolecular: the DNA manipulations here may involve two molecules exchanging parts of their sequences through recombination. The other one, proposed by Ehrenfeucht, Prescott, and Rozenberg, see [11], [27], is intramolecular: here, all manipulations involve one single DNA molecule folding on itself and swapping parts of its sequence through recombination. In the intermolecular model one traditionally attempts to capture both the process of identifying pointers and the process of using pointers by operations that accomplish gene assembly. In the intramolecular model one assumes that the pointer structure of a molecule is known, i.e., the pointers have been already identified. This implies some important differences between the models: e.g., the intramolecular representations of genes contain only pointers, with two occurrences for each pointer, and moreover, processing a pointer implies its removal from the processed string; these properties do not hold in the intermolecular model. Finally, the bulk of the work on the intermolecular model, see [1, 2, 3, 16, 17, 18], is concerned with the computational power of the operations in the sense of computability theory; e.g., it is proved in [18, 19, 20] that the model has the computational power of the Turing machine. On the other hand, research on the intramolecular model, see [4, 5, 6, 8, 9, 10, 12, 13, 14] and especially [7], deals with representations and properties of the gene assembly process (represented by various kinds of reduction systems). We believe that

the two approaches together shed light on the computational nature of gene assembly in ciliates.

In this paper, we take a novel approach on the intermolecular model aiming to compare the assembly strategies predicted by each model. Therefore, we formalize both models in terms of MDS-IES descriptors and describe the gene assembly in terms of pointer reductions. We prove a universality result showing that the assembly power of the two models is the same: any gene that can be assembled in one model can also be assembled in the other. Nevertheless, the assembly strategies and the gene patterns throughout the process are completely different in the two models. Somewhat surprisingly, we show that the two models agree on the final results of the assembly process.

2 The structure of micronuclear genes

We shall now take a formal approach to gene assembly. The central role in this process is played by *pointers*. These are short sequences at the ends of MDSs (i.e., at the border of an MDS and an IES) – the pointer in the end of an MDS M coincides as a nucleotide sequence with the pointer in the beginning of the MDS following M in the macronuclear gene, see [22, 25]. For the purpose of an adequate formal representation, the first (last, resp.) MDS begins (ends, resp.) with a specific marker b (e, resp.). It is enough for our purposes to describe any MDS by the pair of pointers or markers flanking it at its ends. The gene will then be described as a sequence of such pairs interspersed with strings describing the sequence of IES – we thus obtain *MDS-IES descriptors* formally defined in the following. For more details we refer to [7].

For an alphabet Σ and a string u over Σ, we will denote by $[u]$ the circular version of string u – we refer to [7] for a formal definition. Let $\overline{\Sigma} = \{\overline{a} \mid a \in \Sigma\}$ and $u = a_1 a_2 \dots a_n$, $a_i \in \Sigma \cup \overline{\Sigma}$. The *inverse* of u is the string $\overline{u} = \overline{a}_n \dots \overline{a}_2 \overline{a}_1$, where $\overline{\overline{a}} = a$, for all $a \in \Sigma$. The empty string will be denoted by Λ.

Let $\mathcal{M} = \{ b, e, \overline{b}, \overline{e} \}$ denote the set of the *markers* and their inverses. For each index $\kappa \geq 2$, let

$$\Delta_\kappa = \{ 2, 3, \dots, \kappa \} \text{ and } \Pi_\kappa = \Delta_\kappa \cup \overline{\Delta}_\kappa.$$

An element $p \in \Pi_\kappa$ is called a *pointer*. Also let

$$\Gamma_\kappa = \{ (b, e), \} \cup \{ (b, i), (i, e) \mid 2 \leq i \leq \kappa \} \cup \{ (i, j) \mid 2 \leq i < j \leq \kappa \}$$

and $\overline{\Gamma}_\kappa = \{(\overline{\beta}, \overline{\alpha}) \mid (\alpha, \beta) \in \Gamma_\kappa\}$. A string δ over $\Gamma_\kappa \cup \overline{\Gamma}_\kappa$ is called an *MDS descriptor* if

(a) δ has exactly one occurrence from the set $\{b, \overline{b}\}$ and exactly one occurrence from the set $\{e, \overline{e}\}$;

(b) δ has either zero, or two occurrences from $\{p, \overline{p}\}$, for any pointer $p \in \Pi_\kappa$.

Let $\Omega_\kappa = \{I_1, I_2, \dots, I_{\kappa-1}\}$ and $\overline{\Omega}_\kappa = \{\overline{I} \mid I \in \Omega_\kappa\}$. Any string ι over $\Omega_\kappa \cup \overline{\Omega}_\kappa$ is called an *IES-descriptor* if for any $I \in \Omega_\kappa$, ι contains at most one occurrence from $\{I, \overline{I}\}$.

A string γ over $\Gamma_\kappa \cup \overline{\Gamma}_\kappa \cup \Omega_\kappa \cup \overline{\Omega}_\kappa$ is called an *MDS-IES descriptor* if

$$\gamma = \iota_1(p_1, q_1)\iota_2(p_2, q_2) \dots \iota_n(p_n, q_n)\iota_{n+1},$$

where $\iota_1\iota_2\ldots\iota_{n+1}$ is an IES-descriptor, and $(p_1,q_1)\ldots(p_n,q_n)$ is an MDS-descriptor. We say that γ is *assembled* if $\gamma = \iota_1(m,m')\iota_2$ for some IES-descriptors ι_1,ι_2 and $m,m' \in \mathcal{M}$. If $(m,m') = (b,e)$, then we say that γ is assembled in the *orthodox order* and if $(m,m') = (\overline{e},\overline{b})$, then we say that γ is assembled in the *inverted order*.

A circular string $[\gamma]$ is an *(assembled) MDS-IES descriptor* if γ is so.

Example 1. The MDS-IES descriptor associated to the micronuclear *actin I* gene in *S.nova*, shown in Fig. 1, is $M_3I_1M_4I_2M_6I_3M_5I_4M_7I_5M_9I_6\overline{M}_2I_7M_1I_8M_8$. □

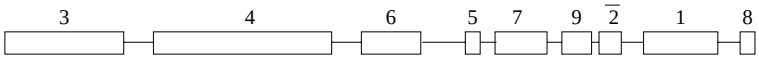


Figure 1: Structure of the micronuclear gene encoding actin protein in the stichotrich *Sterkiella nova*. The nine MDSs are in a scrambled disorder.

3 Two models for gene assembly

We briefly present in this section the intramolecular and the intermolecular models for gene assembly in ciliates. We then formalize both models in terms of pointer reductions and MDS-IES descriptors. For more details we refer to [7, 11, 19, 20, 27].

3.1 The intramolecular model

Three intramolecular operations were postulated in [11] and [27] for gene assembly: ld, hi, and dlad. In each of these operations, a linear DNA molecule containing a specific pattern is folded on itself in such a way that recombination can take place. Operations hi and dlad yield as a result a linear DNA molecule, while ld yields one linear and one circular DNA molecule, see Figs. 2-4. The specific patterns required by each operation are described bellow:

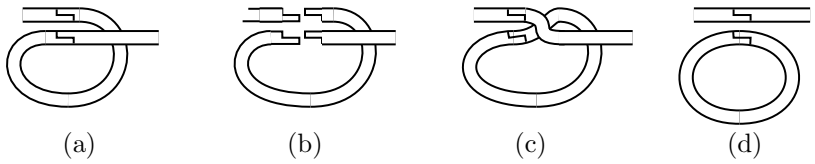


Figure 2: Illustration of the ld molecular operation.

(i) The ld operation is applicable to molecules in which two occurrences (on the same strand) of the same pointer p flank one IES. The molecule is folded so that the two occurrences of p are aligned to guide the recombination, see Fig. 2. As a result, one circular molecule is excised.

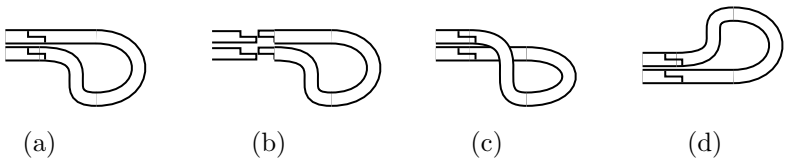


Figure 3: Illustration of the hi molecular operation.

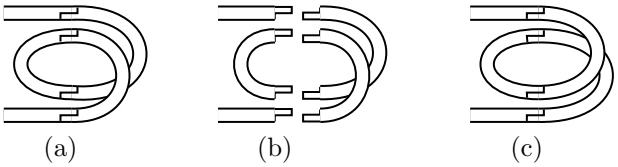


Figure 4: Illustration of the dlad molecular operation.

(ii) The hi operation is applicable to molecules in which a pointer p has two occurrences, of which exactly one is inverted. The folding is done as in Fig. 3 so that the two occurrences of p are aligned to guide the recombination.

(iii) The dlad operation is applicable to molecules in which two pointers p and q have interspersed occurrences (on the same strand): $p - q - p - q$. The folding is done as in Fig. 4 so that the two occurrences of p and q are aligned to guide the double recombination.

Operations ld, hi, and dlad can be formalized in terms of reduction rules ld, hi, and dlad for MDS-IES descriptors as follows:

(1) For each $p \in \Pi_\kappa$, the ld-rule for p is defined by:

$$\underline{\mathsf{ld}}_p(\delta_1(q,p)\iota_1(p,r)\delta_2) = \delta_1(q,r)\delta_2 + [\iota_1]\,,$$
$$\underline{\mathsf{ld}}_p(\iota_1(p,m)\iota_2(m',p)\iota_3) = \iota_1\iota_3 + [(m',m)\iota_2]\,,$$

where $q, r \in \Pi_\kappa \cup \mathcal{M}$, δ_1, δ_2 are MDS-IES descriptors, $\iota_1, \iota_2, \iota_3$ are IES descriptors, and $m, m' \in \mathcal{M}$.

(2) For each $p \in \Pi_\kappa$, the hi-rule for p is defined by:

$$\underline{\mathsf{hi}}_p(\delta_1(p,q)\delta_2(\overline{p},\overline{r})\delta_3) = \delta_1\overline{\delta}_2(\overline{q},\overline{r})\delta_3\,,$$
$$\underline{\mathsf{hi}}_p(\delta_1(q,p)\delta_2(\overline{r},\overline{p})\delta_3) = \delta_1(q,r)\overline{\delta}_2\delta_3\,,$$

where $q, r \in \Pi_\kappa$ and $\delta_1, \delta_2 \in (\Gamma_\kappa \cup \Omega)^{\Phi}$.

(3) For each $p, q \in \Pi_\kappa$, $p \neq q$, the dlad rule for p and q is defined by:

$$\underline{\mathsf{dlad}}_{p,q}(\delta_1(p,r_1)\delta_2(q,r_2)\delta_3(r_3,p)\delta_4(r_4,q)\delta_5) = \delta_1\delta_4(r_4,r_2)\delta_3(r_3,r_1)\delta_2\delta_5\,,$$
$$\underline{\mathsf{dlad}}_{p,q}(\delta_1(p,r_1)\delta_2(r_2,q)\delta_3(r_3,p)\delta_4(q,r_4)\delta_5) = \delta_1\delta_4\delta_3(r_3,r_1)\delta_2(r_2,r_4)\delta_5\,,$$
$$\underline{\mathsf{dlad}}_{p,q}(\delta_1(r_1,p)\delta_2(q,r_2)\delta_3(p,r_3)\delta_4(r_4,q)\delta_5) = \delta_1(r_1,r_3)\delta_4(r_4,r_2)\delta_3\delta_2\delta_5\,,$$
$$\underline{\mathsf{dlad}}_{p,q}(\delta_1(r_1,p)\delta_2(r_2,q)\delta_3(p,r_3)\delta_4(q,r_4)\delta_5) = \delta_1(r_1,r_3)\delta_4\delta_3\delta_2(r_2,r_4)\delta_5\,,$$
$$\underline{\mathsf{dlad}}_{p,q}(\delta_1(p,r_1)\delta_2(q,p)\delta_4(r_4,q)\delta_5) = \delta_1\delta_4(r_4,r_1)\delta_2\delta_5\,,$$
$$\underline{\mathsf{dlad}}_{p,q}(\delta_1(p,q)\delta_3(r_3,p)\delta_4(q,r_4)\delta_5) = \delta_1\delta_4\delta_3(r_3,r_4)\delta_5\,,$$
$$\underline{\mathsf{dlad}}_{p,q}(\delta_1(r_1,p)\delta_2(q,r_2)\delta_3(p,q)\delta_5) = \delta_1(r_1,r_2)\delta_3\delta_2\delta_5\,,$$

where $r_1, r_2, r_3, r_4, r_5 \in \Pi_\kappa$, and $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5 \in (\Gamma_\kappa \cup \Omega)^{\maltese}$.

Note that each operation removes one or two pointers from the MDS-IES descriptor. When assembled (on a linear or on a circular string), the descriptor has no pointers anymore. Thus, the whole process of gene assembly may be viewed as a process of pointer removals.

If a composition φ of ld, hi, and dlad operations is applicable to an MDS-IES descriptor γ, then $\varphi(\gamma)$ is a set of linear and circular MDS-IES descriptors. We say that φ is a *successful reduction* for γ if no pointers occur in any of the descriptors in $\varphi(\gamma)$.

Example 2. Consider the MDS-IES descriptor $\delta = (b,2)I_1(2,3)I_2(4,e)I_3(3,4)$. An assembly strategy for this descriptor in the intramolecular model is the following:

$$\underline{\mathsf{dlad}}_{3,4}(\delta) = (b,2)I_1(2,e)I_3I_2,$$
$$\underline{\mathsf{ld}}_2(\underline{\mathsf{dlad}}_{3,4}(\delta)) = (b,e)I_3I_2 \ + \ [I_1].$$

□

3.2 The intermolecular model

Three operations were postulated in [19] and [20] for gene assembly. One of these operations is intramolecular: it is a sort of a generalized version of the ld operation, while the other two are intermolecular: they involve recombination between two different DNA molecules, linear or circular, see Figs. 5-6. We describe these operations below in terms of pointers, similarly as for the intramolecular model.

(i) In the first operation a DNA molecule containing two occurrences of the same pointer x (on the same strand) is folded so that they get aligned to guide the recombination, see Fig. 5. Note that unlike in ld, the two occurrences of x may have more than just one IES between them.

(ii) The second operation is the inverse of the first one: two DNA molecules, one linear and one circular, each containing one occurrence of a pointer x get aligned so that the two occurrences of x guide the recombination, yielding one linear molecule – see Fig. 5.

(iii) The third operation is somewhat similar to the second one: two *linear* DNA molecules, each containing one occurrence of a pointer x get aligned so that the two occurrences of x guide the recombination, yielding two linear molecules, see Fig. 6.

Note that the three molecular operations in this model are reversible, unlike the operations in the intramolecular model – this is one of the main differences between the two models.

We formalize now this intermolecular model in terms of reduction rules for MDS-IES descriptors. The three operations defined above are modelled by the following reduction rules for MDS-IES descriptors:

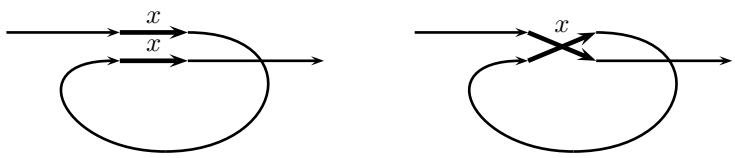


Figure 5: Illustration of the intramolecular operation of the Landweber-Kari model.

Figure 6: Illustration of the intermolecular operation of the Landweber-Kari model.

$$\delta_1(q,p)\delta_2(p,r)\delta_3 \xrightarrow{p} \delta_1(q,r)\delta_3 + [\delta_2], \tag{1}$$

$$\delta_1(p,q)\delta_2(r,p)\delta_3 \xrightarrow{p} \delta_1\delta_3 + [\delta_2(r,q)], \tag{2}$$

$$\delta_1(p,q)\delta_2 + [(r,p)\delta_3] \xrightarrow{p} \delta_1\delta_3(r,q)\delta_2, \tag{3}$$

$$\delta_1(q,p)\delta_2 + [(p,r)\delta_3] \xrightarrow{p} \delta_1(q,r)\delta_3\delta_2, \tag{4}$$

$$\delta_1(p,q)\delta_2 + \delta_3(r,p)\delta_4 \xrightarrow{p} \delta_1\delta_4 + \delta_3(r,q)\delta_2, \tag{5}$$

where $\delta_1, \delta_2, \delta_3 \in (\Gamma_\kappa \cup \overline{\Gamma}_\kappa \cup \Omega_\kappa \cup \overline{\Omega}_\kappa)^*$.

Note that each reduction rule above *removes* one pointer, thus making the whole process irreversible. Although the intermolecular model was specifically intended to be reversible, this restriction helps in unifying the notation for (and the reasoning about) the two models and it suffices for the results presented in this paper.

If a composition φ of the reduction rules (1)-(5) is applicable to an MDS-IES descriptor γ, then $\varphi(\gamma)$ is a set of linear and circular MDS-IES descriptors. We say that φ is a *successful reduction* for γ if no pointers occur in any of the descriptors in $\varphi(\gamma)$.

Example 3. Consider the MDS-IES descriptor $\delta = (b,2)I_1(2,3)I_2(4,e)I_3(3,4)$ of Example 2. An assembly strategy for this descriptor in the intermolecular model is the following:

$$\delta \xrightarrow{3} (b,2)I_1(2,4) \ + \ [I_2(4,e)I_3] \xrightarrow{4} (b,2)I_1(2,e)I_3I_2 \xrightarrow{2} (b,e)I_3I_2 \ + \ [I_1].$$

Note that although the assembly strategy is very different from the one in Example 2, the final result of the assembly, $\{(b,e)I_3I_2, [I_1]\}$ is the same in the two models. □

4 Reduction strategies in the two models

The obvious difficulty with the intermolecular model is that it cannot deal with DNA molecules in which a pointer is inverted – this is the case, e.g., for the *actin I* gene in *S.nova*. Nevertheless, we can show that inverted pointers can be handled in this model, provided the input molecule (or its MDS-IES descriptor) is available in two copies. Moreover, *we consider all linear descriptors modulo inversion*. The first assumption is essentially used in research on the intermolecular model, see [16, 17, 18, 20]. The second assumption is quite natural whenever we model double-stranded DNA molecules. As a matter of fact, we use the two assumptions to conclude that for each input descriptor, both the descriptor and its inversion are available. Then the $\underline{\mathsf{hi}}$-rule can be simulated using the intermolecular rules as follows.

Let $\delta = \delta_1(p,q)\delta_2(\overline{p},\overline{r})\delta_3$ (the other case is treated similarly) be an MDS-IES descriptor to which $\underline{\mathsf{hi}}_p$ is applicable. Therefore, we assume that also $\overline{\delta} = \overline{\delta}_3(r,p)\overline{\delta}_2(\overline{q},\overline{p})\overline{\delta}_1$ is available. Then we obtain

$$\begin{aligned}\delta + \overline{\delta} &\xrightarrow{p} \delta_1\,\overline{\delta}_2\,(\overline{q},\overline{p})\,\overline{\delta}_1 + \overline{\delta}_3\,(r,q)\delta_2\,(\overline{p},\overline{r})\,\delta_3 \\ &\xrightarrow{\overline{p}} \delta_1\overline{\delta}_2(\overline{q},\overline{r})\delta_3 + \overline{\delta}_3(r,q)\delta_2\overline{\delta}_1 = \underline{\mathsf{hi}}_p(\delta) + \overline{\underline{\mathsf{hi}}_p(\delta)}\,.\end{aligned}$$

Note that, having two copies of the initial string available, this rule yields two copies of $\underline{\mathsf{hi}}_p(w)$.

We also observe that the $\underline{\mathsf{ld}}$-rule is a particular case of intermolecular rules (1) and (2), obtained by setting $\delta_2 = \Lambda$. Moreover, the $\underline{\mathsf{dlad}}$-rule can be simulated using intermolecular rules as follows.

Let $\delta = \delta_1(p,r_1)\delta_2(q,r_2)\delta_3(r_3,p)\delta_4(r_4,q)\delta_5)$ be an MDS-IES descriptor to which $\underline{\mathsf{dlad}}_{p,q}$ is applicable – all other cases can be treated similarly. Then

$$\begin{aligned}\delta &\xrightarrow{p} \delta_1\delta_4(r_4,q)\delta_5 + [\delta_2(q,r_2)\delta_3(r_3,r_1)] = \delta_1\delta_4(r_4,q)\delta_5 + [\delta_3(r_3,r_1)\delta_2(q,r_2)] \\ &\xrightarrow{q} \delta_1\delta_4(r_4,r_2)\delta_3(r_3,r_1)\delta_2\delta_5 = \underline{\mathsf{dlad}}_{p,q}(w)\,.\end{aligned}$$

The following results is thus proved.

Theorem 1. *Let δ be an MDS-IES descriptor having a successful reduction in the intramolecular model. If two copies of δ are available, then δ has a successful reduction in the intermolecular model.*

The following universality result has been proved in [8], see [7] for more details.

Theorem 2. *Any MDS-IES descriptor has a successful reduction in the intramolecular model.*

Theorems 1 and 2 give the following universality result for the intermolecular model.

Corollary 3. *Any MDS-IES descriptor available in two copies has a successful reduction in the intermolecular model.*

5 Invariants

In the following two examples we consider the *actin I* gene in *S.nova* and investigate assembly strategies for this gene in the intra- and inter-molecular models.

Example 4. Consider the actin gene in *Sterkiella nova*, see Fig. 1, having the MDS–IES descriptor

$$\delta = (3,4)I_1(4,5)I_2(6,7)I_3(5,6)I_4(7,8)I_5(9,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8(8,9).$$

Consider then an assembly strategy for δ, e.g., $\underline{\mathsf{ld}}_4\ \underline{\mathsf{dlad}}_{5,6}\ \underline{\mathsf{ld}}_7\ \underline{\mathsf{dlad}}_{8,9}\ \underline{\mathsf{hi}}_{\overline{2}}\ \underline{\mathsf{hi}}_3$:

$$\begin{aligned}
&\underline{\mathsf{ld}}_4(\delta) = (3,5)I_2(6,7)I_3(5,6)I_4(7,8)I_5(9,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8(8,9)\ +\ [I_1],\\
&\underline{\mathsf{dlad}}_{5,6}(\underline{\mathsf{ld}}_4(\delta)) = I_0(3,7)I_3I_2I_4(7,8)I_5(9,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8(8,9)\ +\ [I_1],\\
&\underline{\mathsf{ld}}_7(\underline{\mathsf{dlad}}_{5,6}(\underline{\mathsf{ld}}_4(\delta))) = (3,8)I_5(9,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8(8,9)\ +\ [I_1]\ +\ [I_3I_2I_4],\\
&\underline{\mathsf{dlad}}_{8,9}(\underline{\mathsf{ld}}_7(\underline{\mathsf{dlad}}_{5,6}(\underline{\mathsf{ld}}_4(\delta)))) = (3,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8I_5\ +\ [I_1]\ +\ [I_3I_2I_4],\\
&\underline{\mathsf{hi}}_{\overline{2}}(\underline{\mathsf{dlad}}_{8,9}(\underline{\mathsf{ld}}_7(\underline{\mathsf{dlad}}_{5,6}(\underline{\mathsf{ld}}_4(\delta))))) = (3,e)I_6(\overline{3},\overline{b})\overline{I}_7I_8I_5\ +\ [I_1]\ +\ [I_3I_2I_4],\\
&\underline{\mathsf{hi}}_3(\underline{\mathsf{hi}}_{\overline{2}}(\underline{\mathsf{dlad}}_{8,9}(\underline{\mathsf{ld}}_7(\underline{\mathsf{dlad}}_{5,6}(\underline{\mathsf{ld}}_4(\delta)))))) = \overline{I}_6(\overline{e},\overline{b})\overline{I}_7I_8I_5\ +\ [I_1]\ +\ [I_3I_2I_4].
\end{aligned}$$

Thus, the gene is assembled in the inverted order, placed in a linear DNA molecule, with the IES $\overline{I}_6$ preceding it and the sequence of IESs $\overline{I}_7\, I_8\, I_5$ succeeding it. Two circular molecules are also produced: $[I_1]$ and $[I_3I_2I_4]$. □

Example 5. Consider the same actin gene in *Sterkiella nova* with the MDS–IES descriptor

$$\delta = (3,4)I_1(4,5)I_2(6,7)I_3(5,6)I_4(7,8)I_5(9,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8(8,9)\,.$$

Then δ can be assembled in the intermolecular model as follows:

$$\begin{aligned}
\delta\ &\xrightarrow{5}\ (3,4)I_1(4,6)I_4(7,8)I_5(9,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8(8,9) + [I_2(6,7)I_3]\\
&\xrightarrow{8}\ (3,4)I_1(4,6)I_4(7,9) + [I_5(9,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8] + [I_2(6,7)I_3]\\
&\xrightarrow{4}\ (3,6)I_4(7,9) + [I_1] + [I_5(9,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8] + [I_2(6,7)I_3]\\
&\xrightarrow{7}\ (3,6)I_4I_3I_2(6,9) + [I_1] + [I_5(9,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8]\\
&\xrightarrow{6}\ (3,9) + [I_4I_3I_2] + [I_1] + [I_5(9,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8]\\
&\xrightarrow{9}\ (3,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8I_5 + [I_4I_3I_2] + [I_1]\,.
\end{aligned}$$

Since $\overline{\delta}$ is also available, the assembly continues as follows. Here, for a (circular) string τ, we use $2\cdot\tau$ to denote $\tau+\tau$:

$$
\begin{aligned}
\delta + \overline{\delta} \;\; &\rightarrow \ldots \rightarrow (3,e)I_6(\overline{3},\overline{2})I_7(b,2)I_8I_5 + \overline{I}_5\overline{I}_8(\overline{2},\overline{b})\overline{I}_7(2,3)\overline{I}_6(\overline{e},\overline{3}) \\
&\qquad + 2\cdot[I_4I_3I_2] + 2\cdot[I_1] \\
&\xrightarrow{\overline{2}} \quad (3,e)I_6(\overline{3},\overline{b})\overline{I}_7(2,3)\overline{I}_6(\overline{e},\overline{3}) + \overline{I}_5\overline{I}_8\overline{I}_7(b,2)I_8I_5 \\
&\qquad + 2\cdot[I_4I_3I_2] + 2\cdot[I_1] \\
&\xrightarrow{2} \quad (3,e)I_6(\overline{3},\overline{b})\overline{I}_7I_8I_5 + \overline{I}_5\overline{I}_8I_7(b,3)\overline{I}_6(\overline{e},\overline{3}) + 2\cdot[I_4I_3I_2] + 2\cdot[I_1] \\
&\xrightarrow{\overline{3}} \quad (3,e)I_6 + \overline{I}_5\overline{I}_8I_7(b,3)\overline{I}_6(\overline{e},\overline{b})\overline{I}_7I_8I_5 + 2\cdot[I_4I_3I_2] + 2\cdot[I_1] \\
&\xrightarrow{3} \quad \overline{I}_6(\overline{e},\overline{b})\overline{I}_7I_8I_5 + \overline{I}_5\overline{I}_8I_7(b,e) + 2\cdot[I_4I_3I_2] + 2\cdot[I_1] \\
&= \quad 2\cdot(\overline{I}_6(\overline{e},\overline{b})\overline{I}_7I_8I_5 + +[I_1] + [I_3I_2I_4])\,.
\end{aligned}
$$

Note that this intermolecular assembly predicts the same context for the assembled string, the same set of residual strings, and the same linearity of the assembled string as the intramolecular assembly considered in Example 4. □

It is clear from the above two examples, see also Examples 2 and 3, that the two models for gene assembly predict completely different assembly strategies for the same micronuclear gene. As it turns out however, the predicted final result of the assembly, i.e., the linearity of the assembled gene and the exact nucleotide sequences of all excised molecules, is the same in the two models, see [7] for details. The following is a result from [7], see also [10].

Theorem 4. *Let δ be an MDS–IES descriptor. If φ_1 and φ_2 are any two successful assembly strategies for δ, intra- or inter-molecular, then*

(1) if $\varphi_1(\delta)$ is assembled in a linear descriptor, then so is $\varphi_2(\delta)$;

(2) if $\varphi_1(\delta)$ is assembled in a linear descriptor in orthodox order, then so is $\varphi_2(\delta)$;

(3) The sequence of IESs flanking the assembled gene is the same in $\varphi_1(\delta)$ and $\varphi_2(\delta)$;

(4) The sequence of IESs in all excised descriptors is the same in $\varphi_1(\delta)$ and $\varphi_2(\delta)$;

(5) There si an equal number of circular descriptors in $\varphi_1(\delta)$ and $\varphi_2(\delta)$.

Example 6. Consider the MDS–IES descriptor

$$\delta = (\overline{10},\overline{8})I_1(\overline{3},\overline{b})I_2(\overline{5},\overline{3})I_3(10,11)I_4(5,8)I_5(11,e)\,.$$

A successful assembly strategy for δ in the intramolecular model is the following:

$$
\begin{aligned}
&\underline{\mathsf{hi}}_{\overline{10}}(\delta) = \overline{I}_3(3,5)\overline{I}_2(b,3)\overline{I}_1(8,11)I_4(5,8)I_5(11,e), \\
&\underline{\mathsf{dlad}}_{8,11}(\underline{\mathsf{hi}}_{\overline{10}}(\delta)) = \overline{I}_3(3,5)\overline{I}_2(b,3)\overline{I}_1I_5I_4(5,e), \\
&\underline{\mathsf{dlad}}_{3,5}(\underline{\mathsf{dlad}}_{8,11}(\underline{\mathsf{hi}}_{\overline{10}}(\delta))) = \overline{I}_3\overline{I}_1I_5I_4\overline{I}_2(b,e).
\end{aligned}
$$

Thus, δ is always assembled in a linear molecule, and no IES is excised during the assembly process, i.e., no $\underline{\mathsf{ld}}$ is ever applied in a process of assembling δ. Moreover, the assembled descriptor will always be preceded by the IES sequence $\overline{I}_3\overline{I}_1I_5I_4\overline{I}_2$ and followed by the empty IES sequence. □

Example 7. Consider the MDS–IES descriptor

$$\delta = (\overline{10},\overline{8})I_1(\overline{3},\overline{b})I_2(\overline{5},\overline{3})I_3(10,11)I_4(5,8)I_5(11,e)$$

from Example 6. Then δ can be assembled in the intermolecular model as follows:

$$\begin{aligned}\delta &\xrightarrow{\overline{3}} (\overline{10},\overline{8})I_1I_3(10,11)I_4(5,8)I_5(11,e) + [I_2(\overline{5},\overline{b})]\\ &\xrightarrow{11} (\overline{10},\overline{8})I_1I_3(10,e) + [I_4(5,8)I_5] + [I_2(\overline{5},\overline{b})]\,.\end{aligned}$$

Since also $\overline{\delta}$ is available, the assembly continues as follows:

$$\begin{aligned}\delta+\overline{\delta} \;\to\ldots\to\;& (\overline{10},\overline{8})I_1I_3(10,e) + (\overline{e},\overline{10})\overline{I}_3\overline{I}_1(8,10)\\ &+ 2\cdot[I_4(5,8)I_5] + 2\cdot[I_2(\overline{5},\overline{b})]\\ \xrightarrow{\overline{10}}\;& \overline{I}_3\overline{I}_1(8,10) + (\overline{e},\overline{8})I_1I_3(10,e) + 2\cdot[I_4(5,8)I_5] + 2\cdot[I_2(\overline{5},\overline{b})]\\ \xrightarrow{10}\;& \overline{I}_3\overline{I}_1(8,e) + (\overline{e},\overline{8})I_1I_3 + [I_4(5,8)I_5] + [\overline{I}_5(\overline{8},\overline{5})\overline{I}_4] + 2\cdot[I_2(\overline{5},\overline{b})]\\ \xrightarrow{8,\,\overline{8}}\;& \overline{I}_3\overline{I}_1I_5I_4(5,e) + (\overline{e},\overline{5})\overline{I}_4\overline{I}_5I_1I_3 + [(\overline{5},\overline{b})I_2] + [\overline{I}_2(b,5)]\\ \xrightarrow{5,\,\overline{5}}\;& \overline{I}_3\overline{I}_1I_5I_4\overline{I}_2(b,e) + (\overline{e},\overline{b})I_2\overline{I}_4\overline{I}_5I_1I_3\\ =\;& 2\cdot\overline{I}_3\overline{I}_1I_5I_4\overline{I}_2(b,e).\end{aligned}$$

Note that, again, we obtain the same context for the assembled string, the same set of residual strings, and the same linearity of the assembled string as the intramolecular assembly considered in Example 6. □

References

[1] Daley, M., *Computational Modeling of Genetic Processes in Stichotrichous Ciliates*. PhD thesis, University of London, Ontario, Canada (2003)

[2] Daley, M., and Kari, L., Some properties of ciliate bio-operations. *Lecture Notes in Comput. Sci.* **2450** (2003) 116–127

[3] Daley, M., Ibarra, O. H., and Kari, L., Closure propeties and decision questions of some language classes under ciliate bio-operations. *Theoret. Comput. Sci.*, to appear

[4] Ehrenfeucht, A., Harju, T., Petre, I., Prescott, D. M., and Rozenberg, G., Formal systems for gene assembly in ciliates. *Theoret. Comput. Sci.* **292** (2003) 199–219

[5] Ehrenfeucht, A., Harju, T., Petre, I., and Rozenberg, G., Patterns of micronuclear genes in cliates. *Lecture Notes in Comput. Sci.* **2340** (2002) 279–289

[6] Ehrenfeucht, A., Harju, T., Petre, I., and Rozenberg, G., Characterizing the micronuclear gene patterns in ciliates. *Theory of Comput. Syst.* **35** (2002) 501–519

[7] Ehrenfeucht, A., Harju, T., Petre, I., Prescott, D. M., and Rozenberg, G., *Computation in Living Cells: Gene Assembly in Ciliates*, Springer (2003).

[8] Ehrenfeucht, A., Petre, I., Prescott, D. M., and Rozenberg, G., Universal and simple operations for gene assembly in ciliates. In: V. Mitrana and C. Martin-Vide (eds.) *Words, Sequences, Languages: Where Computer Science, Biology and Linguistics Meet*, Kluwer Academic, Dortrecht, (2001) pp. 329–342

[9] Ehrenfeucht, A., Petre, I., Prescott, D. M., and Rozenberg, G., String and graph reduction systems for gene assembly in ciliates. *Math. Structures Comput. Sci.* **12** (2001) 113–134

[10] Ehrenfeucht, A., Petre, I., Prescott, D. M., and Rozenberg, G., Circularity and other invariants of gene assembly in cliates. In: M. Ito, Gh. Paun and S. Yu (eds.) *Words, semigroups, and transductions*, World Scientific, Singapore, (2001) pp. 81–97

[11] Ehrenfeucht, A., Prescott, D. M., and Rozenberg, G., Computational aspects of gene (un)scrambling in ciliates. In: L. F. Landweber, E. Winfree (eds.) *Evolution as Computation*, Springer, Berlin, Heidelberg, New York (2001) pp. 216–256

[12] Harju, T., Petre, I., and Rozenberg, G., Gene assembly in ciliates: molecular operations. In: G.Paun, G. Rozenberg, A.Salomaa (Eds.) *Current Trends in Theoretical Computer Science*, (2004).

[13] Harju, T., Petre, I., and Rozenberg, G., Gene assembly in ciliates: formal frameworks. In: G.Paun, G. Rozenberg, A.Salomaa (Eds.) *Current Trends in Theoretical Computer Science*, (2004).

[14] Harju, T., and Rozenberg, G., Computational processes in living cells: gene assembly in ciliates. *Lecure Notes in Comput. Sci.* **2450** (2003) 1–20

[15] Jahn, C. L., and Klobutcher, L. A., Genome remodeilng in ciliated protozoa. *Ann. Rev. Microbiol.* **56** (2000), 489–520.

[16] Kari, J., and Kari, L. Context free recombinations. In: C. Martin-Vide and V. Mitrana (eds.) *Where Mathematics, Computer Science, Linguistics, and Biology Meet*, Kluwer Academic, Dordrecht, (2000) 361–375

[17] Kari, L., Kari, J., and Landweber, L. F., Reversible molecular computation in ciliates. In: J. Karhumäki, H. Maurer, G. Păun and G. Rozenberg (eds.) *Jewels are Forever*, Springer, Berlin HeidelbergNew York (1999) pp. 353–363

[18] Kari, L., and Landweber, L. F., Computational power of gene rearrangement. In: E. Winfree and D. K. Gifford (eds.) *Proceedings of DNA Bases Computers, V* American Mathematical Society (1999) pp. 207–216

[19] Landweber, L. F., and Kari, L., The evolution of cellular computing: Nature's solution to a computational problem. In: *Proceedings of the 4th DIMACS Meeting on DNA-Based Computers*, Philadelphia, PA (1998) pp. 3–15

[20] Landweber, L. F., and Kari, L., Universal molecular computation in ciliates. In: L. F. Landweber and E. Winfree (eds.) *Evolution as Computation*, Springer, Berlin Heidelberg New York (2002)

[21] Prescott, D. M., Cutting, splicing, reordering, and elimination of DNA sequences in hypotrichous ciliates. *BioEssays* **14** (1992) 317–324

[22] Prescott, D. M., The unusual organization and processing of genomic DNA in hypotrichous ciliates. *Trends in Genet.* **8** (1992) 439–445

[23] Prescott, D. M., The DNA of ciliated protozoa. *Microbiol. Rev.* **58**(2) (1994) 233–267

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

[25] Prescott, D. M., Genome gymnastics: unique modes of DNA evolution and processing in ciliates. *Nat. Rev. Genet.* 1(3) (2000) 191–198

[26] Prescott, D. M., and DuBois, M., Internal eliminated segments (IESs) of Oxytrichidae. *J. Eukariot. Microbiol.* **43** (1996) 432–441

[27] Prescott, D. M., Ehrenfeucht, A., and Rozenberg, G., Molecular operations for DNA processing in hypotrichous ciliates. *Europ. J. Protistology* **37** (2001) 241–260

[28] Prescott, D. M., Ehrenfeucht, A., and Rozenberg, G., Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates. Technical Report 2002-01, LIACS, Leiden University (2002)

[29] Prescott, D. M., and Rozenberg, G., How ciliates manipulate their own DNA – A splendid example of natural computing. *Natural Computing* **1** (2002) 165–183

[30] Prescott, D. M., and Rozenberg, G., Encrypted genes and their reassembly in ciliates. In: M. Amos (ed.) *Cellular Computing*, Oxford University Press, Oxford (2003)

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland
http://www.tucs.fi

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

Åbo Akademi University
- **Department of Computer Science**
- **Institute for Advanced Management Systems Research**

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**