Artiom Alhazov | Ion Petre | Sergey Verlan

A sequence-based analysis of the pointer distribution of ciliate genes

Turku Centre *for* Computer Science

TUCS Technical Report
No 902, September 2008

A sequence-based analysis of the pointer distribution of ciliate genes

Artiom Alhazov
Institute of Mathematics and Computer Science
Academy of Sciences of Moldova
Academiei, 5, MD-2028, Moldova
artiom@math.md

Ion Petre
Academy of Finland and Åbo Akademi University
Turku Center for Computer Science,
FIN-20520 Turku, Finland
ipetre@abo.fi

Sergey Verlan
LACL, Département Informatique, Université Paris 12,
61 av. Général de Gaulle, 94010 Créteil, France
verlan@univ-paris12.fr

TUCS Technical Report
No 902, September 2008

Abstract

It has long been known that the lengths of some of the pointers in the micronuclear ciliate genes are too short to allow for the unambiguous recognition of their coding and noncoding blocks. Many of these pointers have multiple occurrences along the gene, allowing for a very high number of possible divisions into coding and noncoding blocks. We investigate in this paper the pointer distribution of all currently sequenced micronuclear ciliate genes with the goal of identifying what distinguishes the real gene structure among all possible coding/noncoding divisions. We find a surprisingly sharp criterium in the total AT percentage of the IESs of each such division: the real gene has, in most cases, the maximum such percentage among all possible combinations.

Keywords: Ciliates, pointers, sequence analysis, pointer distribution

TUCS Laboratory
Computational Biomodelling Laboratory

1 Introduction

Ciliates are unicellular eukaryotes forming an old and diverse group. They have two types of nuclei: a somatic one, called *macronucleus* (MAC) and a germline one, called *micronucleus* (MIC). Following conjugation, a mitotic copy of the micronucleus develops into a new macronucleus, while the old macronuclei are destroyed. This process involves massive DNA manipulations, including sequence eliminations and rearrangements (inversions and translocations). The process is especially pronounced in an order of ciliates called *Stichotrichs*. This DNA processing is called for by the drastically different genomic organizations in MIC and MAC. Macronuclear genes are continuous sequences of nucleotides, very often placed on their own DNA molecules. The same in the micronucleus is placed on long chromosomes and broken into blocks (called *macronuclear destined sequences*, or *MDS*s), separated by noncoding blocks (called *internally eliminated sequences*, or *IES*s). Moreover, the MDSs are presented in a scrambled order, some of them even being inverted. We refer to [1] for a survey on the topic.

A clue about the mechanism for assembling the MDSs in the orthodox order is given by their structure. It turns out that each MDS ends with a nucleotide sequence that is repeated in the beginning of the MDS that should follow it in the macronuclear gene. These sequences are called *pointers*. In the following, we denote by p_i the pointer that the i-th MDS of the macronuclear gene, say M_i, starts with, for all $i \geq 1$. MDS M_i ends with the pointer p_{i+1} that has an occurrence also in the beginning of MDS M_{i+1}. The sequences in the beginning of the first MDS and at the end of the last MDS are called (beginning and ending, resp.) *markers*.

During gene assembly, the IESs are excised while the MDSs are spliced together on their common pointers to yield the assembled macronuclear gene. It is well understood by now, see [2, 3, 4] that many of the pointers are too short to guarantee unique identification. Indeed, as we also show in this paper, many pointers have multiple occurrences along the micronuclear chromosome. An additional mechanism for the unambiguous identification of all pointers has been proposed in [2] and in [3] and convincingly demonstrated in [4]. The idea is that ciliates would be able to use the old (already assembled) macronuclear gene (or an RNA transcript of it) as a *template*, allowing for unambiguous DNA recognition of whole MDSs rather than just on (short) pointers. For the kinetic details of the proposed template-based recombination mechanisms we refer to [2] and [3].

The motivation of our study is in understanding the difficulty of the pointer identification problem *in the absence of templates*. We consider the problem of identifying the correct occurrences of pointers and so, the correct division into MDSs/IESs, when only the nucleotide sequence of all pointers are known. For all currently sequenced micronuclear ciliate genes, see [5], we consider all possible pairs of occurrences of pointers along the DNA sequence of the micronuclear gene. Some of these combinations lead to alternative divisions into MDSs, while some others do not. Also, some of the alternative MDS sequences may be assembled without losing any MDS, while some others may not. It turns out that in many case, the number of such successful (if alternative) MDS assemblies is huge. We discuss what distinguishes the real assembly among all the other alternatives. This approach is motivated by the following two arguments:

(i) The pointer sequence should be known to ciliates, especially in the absence of a template-based recombination mechanism; otherwise gene assembly

seems impossible.

(ii) We only consider those assemblies that do not omit any of the pointers, mimicking in this way the real assembly.

We only follow one simple criterion in our comparison of all possible combinations: the sum of the A/T-percentage of all the IESs induced by the respective pointer sequence. The results in all case studies we investigated are remarkable: even for pointers as short as two nucleotides, the real assembly is one of very few assemblies with an average A/T-percentage per IES over 80%. This separation is most evident when the shortest pointers (having most occurrences along the chromosome) are fixed on their real positions and only combinations of longer pointers are investigated. Our examples suggest that, as long as the real occurrence of pointers with at most four nucleotides (or even three for unscrambled genes) is known, the real assembly has the maximum A/T-percentage per IES of all possible MDS assemblies. We also discuss in the paper the influence that C/G nucleotides in the pointer sequence have on the result.

Our observations on the A/T-percentage of ciliate IESs should not be taken as a theoretical alternative proposal to template-based recombination in gene assembly. They only establish some unexpected properties of the ciliate genome structure, that may have implications towards the evolution of the mic- and mac- genome organization, rather than towards the kinetic mechanisms of gene assembly.

We introduce first some terminology and notations in Section 2. We then present in Section 3 our methodology and discuss two examples in details in Section 4. We collect in Section 5 the results of several other case studies. Section 6 discusses the conclusions of the paper.

2 Mathematical preliminaries

We introduce in this section some terminology and notations.

For a (possibly infinite) alphabet Σ (whose elements are called *letters*), a *string* over Σ is a finite sequence of letters. We denote by Σ^* the set of all strings over Σ and denote by λ the empty string. For $\alpha \in \Sigma^*$, $\alpha = x_1 x_2 \dots x_n$, $n \geq 0$, $x_i \in \Sigma$, for all $1 \leq i \leq n$, we say that the *length* of α is $|\alpha| = n$. We say that α is a *double occurrence string* if each letter occurring in α has exactly two occurrences in α.

Let $b, e \notin \Sigma$. We say that $\alpha \in (\Sigma \cup \{b, e\}^*$ is an *extended double occurrence string* if it contains exactly one occurrence of b and one of e and the string obtained by deleting b, e from α is a double occurrence string.

For an alphabet $\Sigma = \{a_1, \dots, a_m\}$, consider its *signed copy* $\overline{\Sigma} = \{\overline{a}_1, \dots, \overline{a}_m\}$, where $\overline{\overline{a_i}} = a_i$, for all $1 \leq i \leq n$, and $\Sigma \cap \overline{\Sigma} = \emptyset$. A *signed string* over Σ is any string over the alphabet $\Sigma \cup \overline{\Sigma}$. If $\alpha = x_1 x_2 \dots x_n$, $n \geq 0$, $x_i \in \Sigma \cup \overline{\Sigma}$, for all $1 \leq i \leq n$, we say that $\overline{\alpha} = \overline{x}_n \dots \overline{x}_2 \overline{x}_1$ is its *inverse*. Clearly, the inverse of a signed string over Σ is also a signed string over Σ. The *unsigned copy* of α is $\|\alpha\| = \|x_1\|\|x_2\| \dots \|x_n\|$, where $\|a_j\| = \|\overline{a}_j\| = a_j$, for all $1 \leq j \leq m$. We say that α is a *signed double occurrence string* if $\|\alpha\|$ is a double occurrence string.

We say that $\alpha = x_1 \dots x_n$ as above is an *extended signed double occurrence string* over $\Sigma \cup \{b, e\}$ if $\|\alpha\|$ is an extended double occurrence string. We say that $\beta \in \Sigma^*$ has an *occurrence* in α if

(i) either $\beta = x_r x_{r+1} \dots x_{r+s}$,

(ii) or $\beta = \overline{x}_{r+s} \ldots \overline{x}_{r+1}\overline{x}_r$,

for some $r \geq 1$, $s \geq 0$ with $r + s \leq n$. In the former case we say that β has an occurrence on the direct strand of α at *position* r. In the latter case we say that β has an occurrence on the inverse strand of α at *position* $-r$ (which is a negative integer). Clearly, each occurrence of β is uniquely identified by its position.

Let α, β, γ be signed strings over Σ and i, j two integers with $|j| > |i|$. Assume that β has an occurrence within α at position i and γ has an occurrence within α at position j. We say that these two occurrence *overlap* if $|j| - |i| < |\beta|$. Note that in this definition γ may coincide with β.

We represent genes as sequences of nucleotides, i.e., as strings over alphabet $N = \{a, c, g, t\}$, where $\overline{a} = t$, $\overline{t} = a$, $\overline{c} = g$, $\overline{g} = c$. In the case of ciliate micronuclear genes, we may also represent the genes by the sequences of their MDSs and IESs. E.g., $H\ M_2\ I\ \overline{M}_1\ J$ would represent a micronuclear gene with MDSs M_1, M_2, with M_1 being inverted and placed after M_2, and IESs H, I, J separating them. As discussed in Section 1, the MDSs have a special structure and each MDS may be uniquely identified by its beginning pointer or marker. The general convention is to denote by M_1 the MDS starting with the beginning marker (and ending with pointer p_2) and by M_i the MDS starting with pointer p_i of the gene (and ending either with p_{i+1} or, in the case of the last MDS, with the ending marker). A similar convention can be made also for denoting the IESs. E.g., we may denote by I_k the IES having the k-th pointer (even if inverted) at its left extremity. We denote by I_1 the IES having b at its left extremity and by I_{n+1} the IES having e at its left extremity. The very first IES of the micronuclear gene, occurring before all the MDS has no marker or pointer at its left extremity. For its notation we use the same convention as above, using however the pointer or marker at its right extremity. We call *MDS/IES sequence* any such string. E.g., the MDS/IES sequence associated to the gene above is $I_2 M_2\ I_3\ \overline{M}_1\ I_1$.

Given the MDS/IES sequence α of a micronuclear gene, the sequence of its pointers and markers, say $\phi(\alpha)$, is easy to deduce, as discussed above. Consider now the inverse problem: we are given a signed double occurrence string of pointers and markers (where we denote by b and e the two occurrences of the marker), determine whether it corresponds to an MDS/IES sequence. For reasons that become apparent in Section 3, we take in fact a more general formulation of the problem, where the pointers may be relabeled (e.g., pointer p_2 may be the beginning pointer of the 5-th MDS). We also allow that the orientation of some of the pointers may be changed. We call such strings realizable and we define them formally in the following. We first give an example.

Example 1. The string $b\ 2\ 2\ 3\ 3\ e$ clearly corresponds to the MDS/IES sequence $I_1 M_1 I_2 M_2 I_3 M_3 I_4$. The string $b\ \overline{3}\ \overline{3}\ 2\ 2\ e$ does not correspond to any such sequence. Nevertheless, a suitable transformation of the string, where 2 is relabeled as 3 and 3 is relabeled as $\overline{2}$ (and by consequence, $\overline{3}$ is relabeled as 2) yields the string $b\ 2\ 2\ 3\ 3\ e$.

Definition 1. *Let $\Delta = \{p_1, \ldots, p_n\}$ be an alphabet of pointers and b, e other two distinct letters. We say that an extended signed double occurrence string u over $\Delta \cup \{b, e\}$ is realistic if there exists an MDS/IES sequence α such that $\phi(\alpha) = u$. We say that α is its induced MDS/IES sequence.*

We say that u is realizable if there exists a string morphism $\psi : \Delta \to \Delta \cup \overline{\Delta}$ such that $\psi(u)$ is realistic.

The notion of realistic string captures the notion of sequences of pointers and markers coming directly from MDS/IES sequences. The notion of realizable strings captures the notion of sequences of pointers and markers that are coming from MDS/IES sequence but where the notation of pointers did not respect the convention we described above for denoting MDSs and IESs. Such a situation may be encountered for example in the case where one would know the nucleotide sequence of all pointers, but would not know which pointer each of them is, neither the strand from which they were read.

It is not difficult to see that the string morphism ψ in Definition 1 is unique, since b and e are fixed. Consequently, the division of a realizable string u into MDSs and IESs is unique: the MDSs are those of $\psi(u)$, while the IESs are the blocks separating the MDSs. To denote the IESs of u, we use the convention described above, based on the pointer at the left extremity of each IES (the right extremity for the very first IES).

Example 2. As noted in Example 1, $u = b\ \overline{3}\ \overline{3}\ 2\ 2\ e$ is realizable. Indeed, if ψ is defined as $\psi(2) = 3$, $\psi(3) = \overline{2}$, then $\psi(u) = b\ 2\ 2\ 3\ 3\ e$, having the MDS sequence $M_1\ M_2\ M_3$. The induced MDS/IES sequence of $\psi(u)$ and of u is $I_1\ M_1\ I_3\ M_2\ I_2\ M_3\ I_4$.

3 The approach

We consider all ciliate genes from [5] for which the nucleotide sequences of all pointers are known. We include also the DNA polymerase Alpha gene in P.Weissei, for which only the last pointer is not known. In this case we have replaced the unknown pointer by the end marker. The list of genes considered in our study is summarized in Table 1.

Organism	Gene	MIC (bp)	MAC (bp)	Ptrs
S. Nova	Actin I	2374	1604	8
S. Histriomuscorum	Actin I	2115	1558	9
S. Nova	Alpha Telomere Binding Protein	2700	2217	13
E. octocarinatus	Aminoacyl-tRNA Synthetase Cofactor	2516	1517	2
S. Histriomuscorum	Beta Telomere Binding Protein	2336	1858	6
S. Mytilus	Alpha Telomere Binding Protein	2686	2141	13
E. octocarinatus	cAMP-Dependent Protein Kinase Regulatory Subunit	1409	1398	1
E. octocarinatus	Gamma-Tubulin 2	2124	1633	1
S. Nova	Beta Telomere Binding Protein	1839	1790	3
S. mytilus	Beta Telomere Binding Protein	1739	1738	2
S. Nova	C2	1200	737	3
S. Nova	R1	2035	1029	5
P. Weissei	DNA Polymerase Alpha gene	6930	4746	47

Table 1: Summary of the genes analyzed in our study.

In our analysis we will consider the scenario where the assembly machinery is only aware of the pointer sequence, but not of their occurrences, nor of the strands of their occurrences along the micronuclear gene. In particular, the machinery does not know which of the pointers should come first, which second, etc. To simplify the analysis, we assume that the the sequence and the exact position of the beginning and end markers are known. The machinery must determine what are the real occurrences of the pointers (two occurrences for each

of them) and assemble the MDSs that are induced by those pointer occurrences. The difficulty in this setup comes when pointers have multiple occurrences. Then one has to distinguish among a large number of possible combinations of pointer occurrences, which induce very different MDS and IES blocks.

Example 3. Consider a hypothetical gene sequence of the form

$$b\ \alpha_1\ p\ \alpha_2\ q\ \alpha_3\ r\ \alpha_4\ p\ \alpha_5\ q\ \alpha_6\ r\ \alpha_7\ p\ \alpha_8\ q\ \alpha_9\ e,$$

where p, q, r denote the nucleotide sequences of the pointers of the gene and b, e those of its (beginning and ending, resp.) markers and $\alpha_1, \ldots, \alpha_7$ are arbitrary sequences. Assuming that the nucleotide sequences of p, q, r are known to the assembly machinery, one must still distinguish among several possible MDS decompositions, leading to very different assembly results. Here are three possible MDS decompositions, where we indicate by parenthesis the extremities of each MDS:

- $(b\ \alpha_1\ p)\ \alpha_2\ (q\ \alpha_3\ r)\ \alpha_4\ (p\ \alpha_5\ q)\ \alpha_6\ (r\ \alpha_7\ p\ \alpha_8\ q\ \alpha_9\ e)$, leading to the assembled gene $(b\ \alpha_1\ p\ \alpha_5\ q\ \alpha_3\ r\ \alpha_7\ p\ \alpha_8\ q\ \alpha_9\ e)$;
- $(b\ \alpha_1\ p\ \alpha_2\ q)\ \alpha_3\ (r\ \alpha_4\ p)\ \alpha_5\ (q\ \alpha_6\ r)\ \alpha_7\ (p\ \alpha_8\ q\ \alpha_9\ e)$, leading to the assembled gene $(b\ \alpha_1\ p\ \alpha_2\ q\ \alpha_6\ r\ \alpha_4\ p\ \alpha_8\ q\ \alpha_9\ e)$;
- $(b\ \alpha_1\ p\ \alpha_2\ q\ \alpha_3\ r)\ \alpha_4\ (p\ \alpha_5\ q)\ \alpha_6\ (r\ \alpha_7\ p)\ \alpha_8\ (q\ \alpha_9\ e)$, leading to the assembled gene $(b\ \alpha_1\ p\ \alpha_2\ q\ \alpha_3\ r\ \alpha_7\ p\ \alpha_5\ q\ \alpha_9\ e)$.

Note that the assembled genes are different in these three cases.

An additional difficulty is noted in the following example, showing that some sequences of pointers and markers may not induce at all a division of the gene into MDSs.

Example 4. Consider a gene of the form $b \ldots p \ldots \overline{p} \ldots e$, where p denotes the nucleotide sequence of the gene's only pointer and b, e those of its (beginning and ending, resp.) markers. Note that these two occurrences of p do not induce a division of the gene into MDSs. Such a division should start with the beginning marker b and end with one of the two occurrences of p. If it ends with $\overline{p}$, then there is no other 'free' occurrence of p to use in the following MDS, that should end with e. If the first MDS is $(b \ldots p)$, then the second MDS should start with the second occurrence of p. Since that occurrence is inverted, then the whole MDS should be inverted and so, the other pointer or marker of the second MDS should be found in-between p and $\overline{p}$. No such pointer or marker exists in our example.

For all genes in Table 1, we use the following algorithmic procedure:

1. Consider the nucleotide sequences of all the pointers of the gene. For every pointer sequence, find all its occurrences on both strands of the gene.
2. Consider all possible combinations of non-overlapping pointer occurrences having exactly two occurrences of each pointer. Each combination yields an extended signed double occurrence string of pointers and markers.
3. For all realizable such strings, output the sum of the AT percentage of all its induced IESs.

We implemented the algorithm above in *Perl*. We present in the following the resulting AT-percentage data for all genes in Table 1.

4 Two examples

In this section we discuss in details our analysis of the pointer distribution in two genes: Actin I and Beta Telomere Binding Protein, both in S. Nova.

4.1 Actin I in S.nova

The next table summarizes the nucleotide sequence of the pointers of this gene and the number of their occurrences throughout the gene:

No	Sequence	Occurrences	No	Sequence	Occurrences
2	cttactacacat	2	6	agcccc	3
3	cggagtcgtcaag	2	7	caaaactcta	2
4	aatc	17	8	cctttgggttga	2
5	ctcccaagtccat	2	9	aggttgaatga	2

Table 2: The distribution of pointers in the micronuclear gene actin I in S.nova.

This gives us 408 possible combinations of pairs of pointers. After Step 4 of the algorithm there remain only 7 variants that lead to a correct gene assembly. Moreover, it turns out that the position of pointer P6 cannot be varied because all other combinations lead to an incorrect gene assembly.

Next we computed the AT-percentage for all IES sequences and we found that only IES 4 is different. The diagram below contains corresponding values.

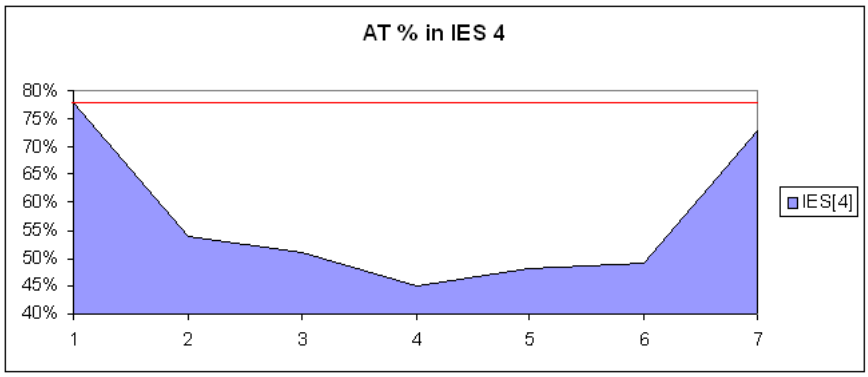


Figure 1: The micronuclear actin I gene in S.nova. The AT percentage of IES 4 in all combinations of all pointer occurrences. The red line indicates the percentage of the real pointer distribution.

As one can see, the real position has the highest AT-percentage (at 78%). We also remark that 10 of 11 IESs for this gene have an AT-percentage higher than 70%.

4.2 Beta telomere binding protein in S.nova

Table 4.2 summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence. This gives us 1026 possible combinations of pairs of pointers. After Step 4 of the algorithm there remain only 42 variants that lead to a correct gene assembly.

No	Sequence	Occurrences	No	Sequence	Occurrences
2	gtcca	4	4	agtc	19
3	taaagt	2			

Table 3: The distribution of pointers in the micronuclear gene encoding for the beta telomere binding protein in S.Nova.

Next we compute the AT-percentage for all IES sequences and we found that only IES 4 and IES 2 are different. The diagram below contains the sum of corresponding values.

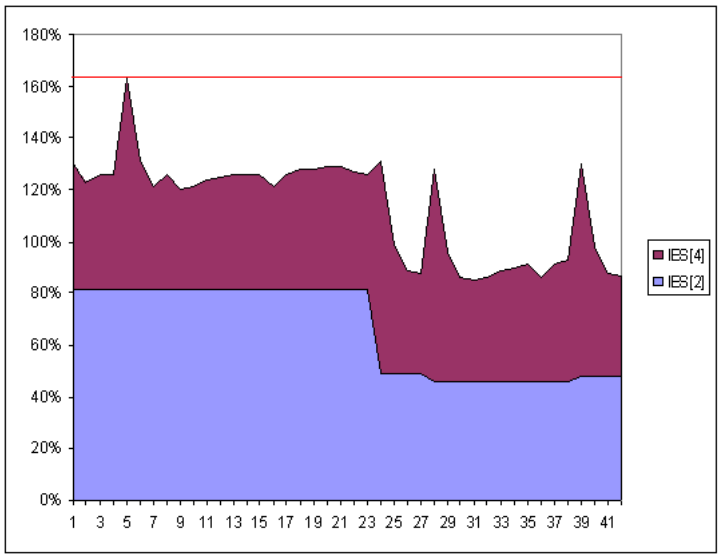


Figure 2: The micronuclear gene encoding for the beta telomere binding protein in S.nova. The AT percentage in IESs 2 and 4. The red line indicates the percentage of the real pointer distribution.

As one can see, the real position has the highest sum of AT-percentage (at 163%). We note that all other IES have an AT-percentage higher than 75%.

5 The other case studies

In this section we present all other results.

5.1 Gamma-tubulin 2 in E.octocarinatus

The next table summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence:

No	Sequence	Occurrences
2	gatatt	5

Table 4: The distribution of pointers in the micronuclear gene encoding for the gamma-tubulin 2 gene in E.octocarinatus.

This gives us 10 possible combinations of pairs of pointers. After Step 4 of the algorithm there remain only 3 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 2 is different. The diagram below contains corresponding values. The red line indicates the AT-percentage of the real pointer distribution.

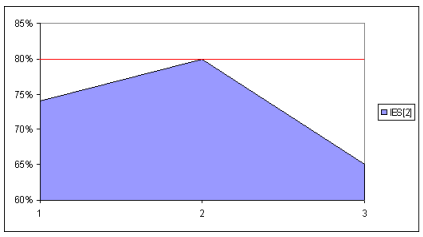


Figure 3: The micronuclear gamma-tubulin 2 gene in E.octocarinatus. The AT percentage in IES 2 in all combinations of pointer occurrences. The red line indicates the percentage of the real pointer distribution.

5.2 The cAMP-dependent protein kinase regulatory subunit in E.octocarinatus

The next table summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence:

No	Sequence	Occurrences
2	taca	13

Table 5: The distribution of pointers in the micronuclear gene encoding for the cAMP-dependent protein kinase regulatory subunit in E.octocarinatus.

This gives us 78 possible combinations of pairs of pointers. After Step 4 of the algorithm there remain only 18 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 2 is different. The diagram below contains corresponding values. The red line indicates the AT-percentage of the real pointer distribution.

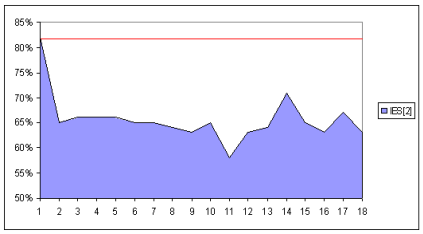


Figure 4: The micronuclear cAMP-dependent protein kinase regulatory subunit in E.octocarinatus. The AT percentage in IES 2 of all combinations of pointer occurrences. The red line indicates the percentage of the real pointer distribution.

5.3 Beta telomere binding protein in S.mytilus

The next table summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence:

No	Sequence	Occurrences	No	Sequence	Occurrences
2	atgtt	5	3	gaaaga	14

Table 6: The distribution of pointers in the micronuclear gene encoding for the beta telomere binding protein in S.mytilus.

This gives us 60 possible combinations of pairs of pointers. After Step 4 of the algorithm there remain only 8 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 2 and IES 3 are different. The diagram below contains corresponding values. The red line indicates the AT-percentage of the real pointer distribution.

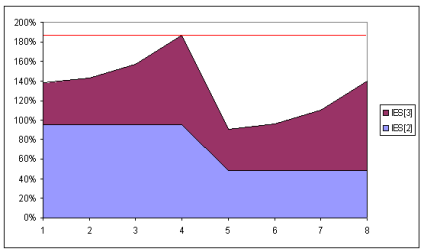


Figure 5: The micronuclear gene encoding for the beta telomere binding protein in S.mytilus. The AT percentage in IESs 2 and 3 of all combinations of pointer occurrences. The red line indicates the percentage of the real pointer distribution.

5.4 Actin I in S.histriomuscorum

The next table summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence:

No	Sequence	Occurrences	No	Sequence	Occurrences
2	agaccaacaaa	2	7	tgaggaatcaaat	2
3	aaggctggtttc	2	8	gggttgaatga	2
4	tctc	28	9	aggttgaatga	2
5	agctcccaagtca	2	10	caaaaat	3
6	tattgcca	2			

Table 7: The distribution of pointers in the micronuclear gene encoding for actin I in S.histriomuscorum.

This gives us 1134 possible combinations of pairs of pointers. After Step 4 of the algorithm there remain only 58 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 4 is different. The diagram below contains corresponding values. The red line indicates the AT-percentage of the real pointer distribution.

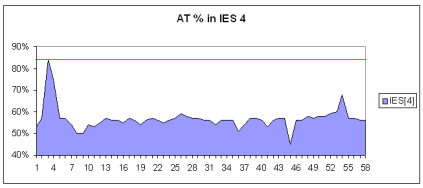


Figure 6: The micronuclear gene encoding for actin I in S.histriomuscorum. The AT percentage in IES 4 of all combinations of pointer occurrences. The red line indicates the percentage of the real pointer distribution.

5.5 Beta telomere binding protein in S.histriomuscorum

The next table summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence:

No	Sequence	Occurrences	No	Sequence	Occurrences
2	cagta	3	5	act	37
3	acatttc	3	6	actgct	2
4	actc	19	7	agt	78

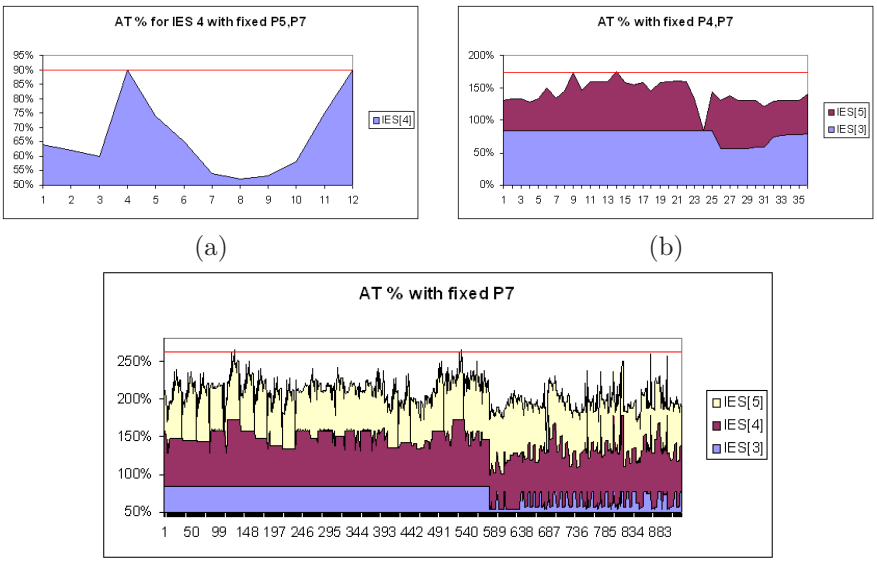
Table 8: The distribution of pointers in the micronuclear gene encoding for the beta telomere binding protein in S.histriomuscorum.

This gives us 3,077,996,922 possible combinations of pairs of pointers. Due to combinatorial problems we did only a subset of them. We firstly varied positions of pointers P2, P3 and P4, while keeping the position of P5 and P7 to the correct one (1539 combinations). After that we varied positions of pointers P2, P3 and P5, while keeping the position of P4 and P7 to the correct one (5994 combinations). Finally, we varied the positions of pointers P2, P3, P4 and P5, while keeping the position of P7 to the correct one (1024974 combinations).

We first analyze the distribution of pointers P2, P3 and P4 and consider all combinations of their occurrences. All the other pointers remain fixed on their real positions. After Step 4 of the algorithm there remain only 12 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 4 is different, see Figure 7 (a).

Consider now all combinations of occurrences of pointers P2, P3, and P5, while keeping all other pointers fixed on their real positions. After Step 4 of the algorithm there remain only 36 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 3 and IES 5 are different, see Figure 7 (b). There is only one value (at 173%) greater than the real position (171%).

Finally, consider the combinations of all occurrences of P2, P3, P4, P5, with all other pointers fixed on their real positions. After Step 4 of the algorithm there remain only 923 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 3, IES 4 and IES 5 are different, see Figure 7 (c). There are 3 combinations greater or equal to the real one at 263% (the maximal difference is 2% and the average difference is 1,3%).



(a) (b)

(c)

Figure 7: The beta telomere binding protein in S.histriomuscorum. The AT percentage in (a) IES 4 of all combinations of occurrences of P2, P3, and P4; (b) IESs 3 and 5 of all combinations of occurrences of P2, P3, and P5; (c) IESs 3, 4, 5 of all combinations of occurrences of P2, P3, P4, and P5. The red line indicates the percentage of the real pointer distribution.

5.6 Alpha telomere binding protein in S.nova

The next table summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence:

No	Sequence	Occurrences	No	Sequence	Occurrences
2	gaaggcgctgc	2	9	aaggac	5
3	gccaccctc	2	10	aagtgttct	2
4	tcatccaca	2	11	agaact	4
5	agagctaccctc	2	12	gaatcagatcagccactta	2
6	tcaagcaag	2	13	cccaa	6
7	ttgagaagaacga	2	14	act	88
8	agaacctga	2			

Table 9: The distribution of pointers in the micronuclear gene encoding for the alpha telomere binding protein in S.nova.

This gives us 3445200 possible combinations of pairs of pointers. Due to combinatorial problems we initially did only a subset of them. We firstly varied positions of pointers P9, P11 and P13, while keeping the position of P14 to the correct one (900 combinations). After that we varied only the positions of P14 (3828 combinations). Finally we were able to check all combinations.

Consider all combinations of occurrences of P9, P11 and P13, while keeping all other pointers fixed on their occurrences. After Step 4 of the algorithm there remain only 2 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that most IES are different, see Figure 8 (a).

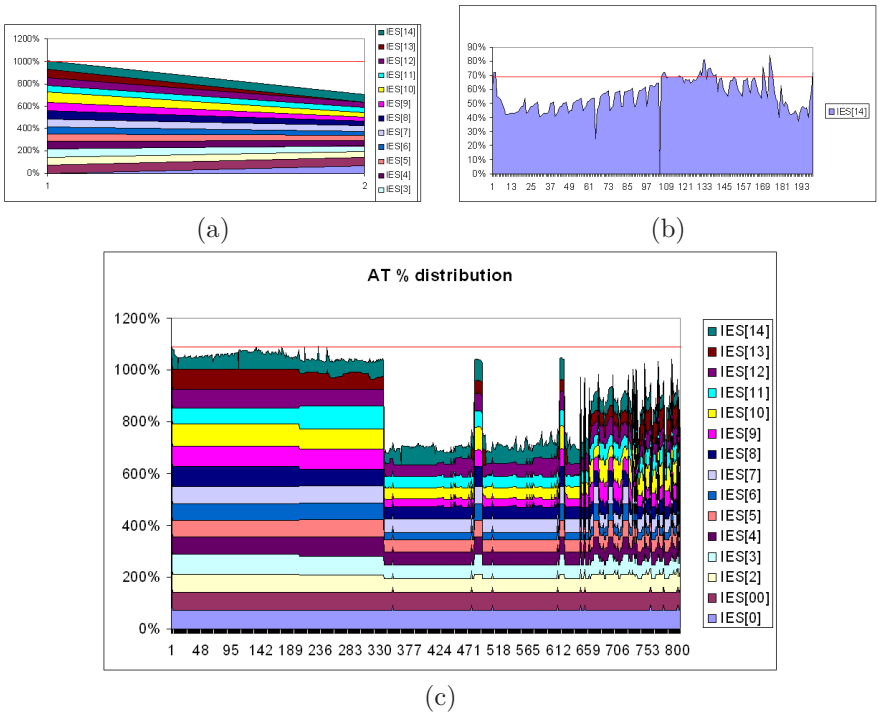


Figure 8: The alpha telomere binding protein in S.nova. The total AT percentage in (a) the IESs of all combinations of occurrences of P9, P11, P13; (b) IES 14 of all combinations of occurrences of P14; (c) the IESs of all combinations of occurrences of all pointers. The red line indicates the percentage of the real pointer distribution.

We consider now all possible occurrences of P14. After Step 4 of the algorithm there remain only 201 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 14 is different, see Figure 8 (b).

Consider now all combination of occurrences of all pointers. After Step 4 of the algorithm there remain only 805 (out of 3445200) variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that most IES are different, see Figure 8 (c). There are 37 positions (out of 805) that are greater than or equal to the real one (the maximal difference is 19%, the average difference is 4%).

5.7 Alpha telomere binding protein in S.mytilus

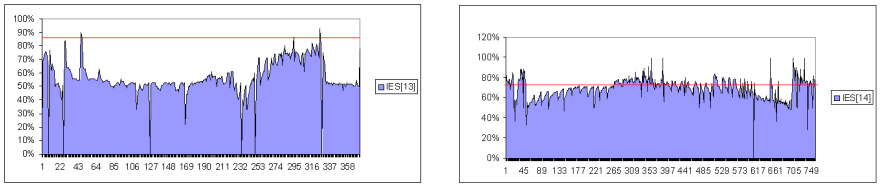
Table 10 summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence. This gives us 232,240,638 possible combinations of pairs of pointers. Due to combinatorial problems we did only a subset of them. We firstly varied positions of pointers P11, and P13, while keeping the position of P14 correct one (33078 combinations). After that we varied positions of pointers P11 and P14, while keeping the position of P13 to the correct one (21063 combinations).

Consider all combinations of all occurrences of P11 and P13. After Step 4 of the algorithm, there remain only 372 variants that lead to a correct gene

No	Sequence	Occurrences	No	Sequence	Occurrences
2	taaagacggcgcaccaaag	2	9	ctcaagttgaa	2
3	tagtcttat	2	10	aagttttct	2
4	gaatcggaga	2	11	ggagaag	3
5	cagagctactctca	2	12	atcagctacttat	2
6	ccattcg	2	13	aag	149
7	ttgagaagagcga	2	14	aat	119
8	ccttctcca	2			

Table 10: The distribution of pointers in the micronuclear gene encoding for the alpha telomere binding protein in S.mytilus. The red line indicates the percentage of the real pointer distribution.

assembly. We computed the AT-percentage for all IES sequences and we found that only IES 13 is different. The diagram below contains corresponding values.



(a) (b)

Figure 9: The alpha telomere binding protein in S.mytilus. The total AT percentage in (a) IES 13 of all combinations of occurrences of P11 and P13; (b) IES 14 of all combinations of occurrences of P11 and P14. The red line indicates the percentage of the real pointer distribution.

There are 3 combinations greater or equal to the real one at 86% (with the maximal difference 7% and average difference 4%). We note that if P13 and P14 are fixed to correct values, then there is only one possible assembly, the real one.

We consider now all occurrences of pointers P11 and P14, while fixing all other pointers on their real positions. After Step 4 of the algorithm there remain only 758 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 14 is different, see Figure 9 (b). There are 280 combinations greater or equal to the real one at 73% (with the maximal difference 17% and average difference 4,6%).

5.8 DNA polymerase alpha gene in P.weissei

Table 11 summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence. This gives us about 2.4×10^{53} possible combinations of pairs of pointers. Due to combinatorial problems we did only a subset of them. We varied positions of pointers having the size greater than or equal to 8, firstly P4, P8, P12, P25, P31, P36, P38 and P39 (244944 combinations), then we varied positions of P4, P8, P10, P12, P25, P36, P38 and P39 (682344 combinations).

Consider first the distribution of P4, P8, P12, P25, P31, P36, P38 and P39. After Step 4 of the algorithm there remain 2 (out of 244944) variants that lead to a correct gene assembly. We computed the AT-percentage for all

No	Sequence	Occurrences	No	Sequence	Occurrences
2	atgc	51	26	aatggtttta	2
3	aaat	542	27	ttatgtggt	2
4	ttaacatt	2	28	attccaaata	2
5	gaaa	104	29	gaggatcatagt	2
6	aagcag	7	30	aagata	16
7	aaagcaaca	2	31	aaaattaa	8
8	gattataa	3	32	aagaga	14
9	attatcttt	2	33	ttgctga	3
10	aaaataat	13	34	tattatgattaat	2
11	aatatgtct	2	35	gagtttttaa	2
12	agaaatata	3	36	ttaaagta	4
13	aaat	542	37	gtaatta	5
14	actctta	5	38	ataaaatga	3
15	aatcataataagtta	2	39	aataacttt	3
16	tagcat	9	40	aaagtgaagct	2
17	aaatgaagtat	2	41	agtcaacaatt	2
18	aaaca	27	42	tgatatg	6
19	aatgctt	3	43	tttgta	8
20	aaactaaa	2	44	tgatttt	4
21	aaaaacttg	2	45	agagggt	2
22	aagttactctt	2	46	aaattat	16
23	aaaataaca	2	47	ta	1648
24	tatgatatcat	2	48	aat	704
25	atttgatt	4			

Table 11: The distribution of pointers in the micronuclear gene encoding for the DNA polymerase alpha gene in P.weissei.

IES sequences and we found that only IES 32 is different. The diagram below contains corresponding values.

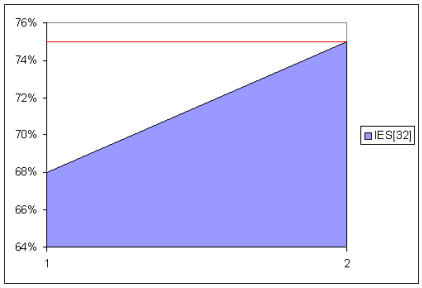


Figure 10: The micronuclear gene encoding for the DNA polymerase alpha gene in P.weissei. The total AT percentage in IES 32 of all combinations of occurrences of P4, P8, P12, P25, P31, P36, P38 and P39. The red line indicates the percentage of the real pointer distribution.

When the distribution of P4, P8, P12, P25, P31, P36, P38 and P39 is considered, after step 4 of the algorithm only the real assembly remains out of 682344 possible combinations.

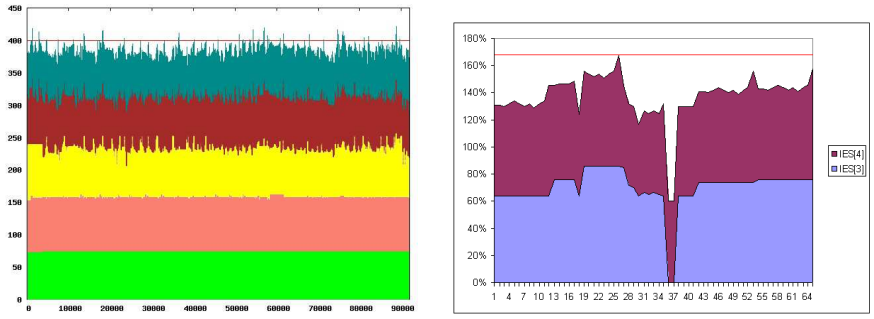
5.9 C2 in S.nova

The next table summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence:

No	Sequence	Occurrences	No	Sequence	Occurrences
2	agt	48	4	agca	11
3	tgag	12			

Table 12: The distribution of pointers in the micronuclear gene encoding for the C2 micronuclear gene in S.nova

This gives us 4,094,640 possible combinations of pairs of pointers. After Step 4 of the algorithm there remain 91,777 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that most IES are different. The diagram below contains corresponding values.



(a) (b)

Figure 11: The micronuclear gene C2 in S.nova. The total AT percentage in (a) all IESs of all combinations of pointer occurrences and (b) all IESs of all combinations of pointer occurrences except P2. The red line indicates the percentage of the real pointer distribution.

There are 147 assemblies (out of 91,777) having the sum of AT-percentage for all IESs greater or equal to the real one at 400% (with the maximal difference 21% and average difference 4,3%).

If the value of P2 is fixed, then the resulting assembly has the maximal AT-percentage.

5.10 R1 in S.nova

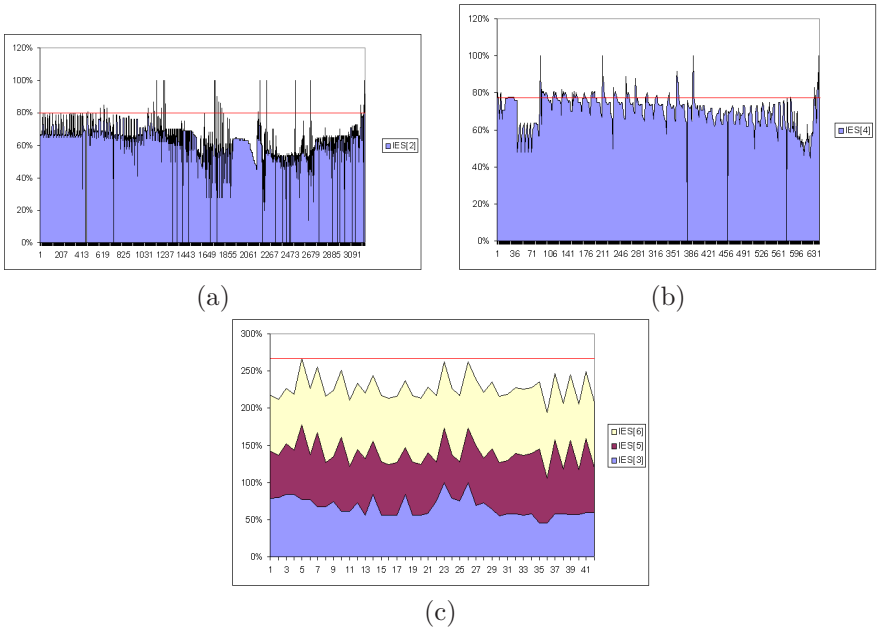
The next table summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence:

No	Sequence	Occurrences	No	Sequence	Occurrences
2	aa	510	5	atttat	13
3	tagc	12	6	atcact	3
4	aat	167			

Table 13: The distribution of pointers in the micronuclear gene encoding for the R1 micronuclear gene in S.nova

This gives us 27,785,122,716,780 possible combinations of pairs of pointers. Due to combinatorial problems we did only a subset of them. We firstly varied positions of P2 only (129795 combinations), then we varied positions of P4 only (13861 combinations) and finally we varied P3, P5 and P6 (15444 combinations).

Consider first all possible occurrences of P2. After Step 4 of the algorithm there remain only 3213 (out of 129795) variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 2 is different, see Figure 12 (a).



(a) (b)

(c)

Figure 12: The micronuclear gene R1 in S.nova. The total AT percentage in (a) IES 2 of all combinations of pointer occurrences of P2; (b) IES 4 of all combinations of pointer occurrences of P4; (c) IESs 3,5, 6 of all combinations of pointer occurrences of P3, P5, P6. The red line indicates the percentage of the real pointer distribution.

For the distribution of P4, after Step 4 of the algorithm there remain 642 (out of 13861) variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 4 is different, see Figure 12 (b).

Finally, for the distribution of P3, P5, P6, after Step 4 of the algorithm there remain 42 (out of 15444) variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 3, IES 5 and IES 6 are different, see Figure 12 (c).

5.11 The aminoacyl-tRNA synthetase cofactor in E.octocarinatus

The next table summarizes the pointers present in this gene and the number of their occurrences throughout the gene sequence:

No	Sequence	Occurrences	No	Sequence	Occurrences
2	ttactga	3	3	ta	467

Table 14: The distribution of pointers in the micronuclear gene encoding for the micronuclear gene encoding for the aminoacyl-tRNA synthetase cofactor in E.octocarinatus.

This gives us 326433 possible combinations of pairs of pointers. After Step 4 of the algorithm there remain 15576 variants that lead to a correct gene assembly. We computed the AT-percentage for all IES sequences and we found that only IES 4 and IES 5 are different. The diagram below contains corresponding values. The red line indicates the AT-percentage of the real pointer distribution.

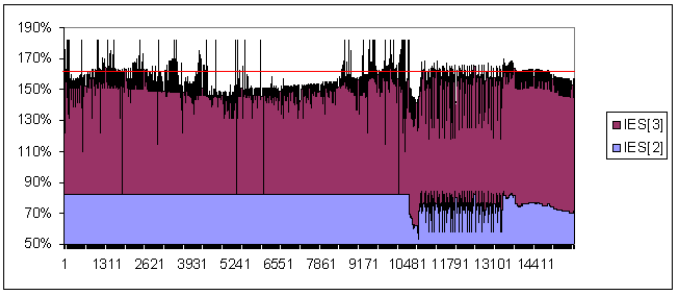


Figure 13: The micronuclear gene encoding for the aminoacyl-tRNA synthetase cofactor in E.octocarinatus. The total AT percentage in IESs 4, 5 of all combinations of all pointer occurrences. The red line indicates the percentage of the real pointer distribution.

We note that if P3 is fixed to correct values, then there is only one possible assembly: the real one.

Although the separation is less clear than in previous examples the result is still remarkable given that one of the pointers is only 2 bp long. Note that the sequence ta would appear just by chance in average every 16 positions in a random DNA sequence and every 4 positions in AT-reach non-coding sequences.

6 Conclusions

We investigated in this paper the difficulty of the pointer identification problem in gene assembly, in the absence of a template-based mechanism as in [2, 3]. We showed that the multiple occurrences of pointer sequences lead to a high number of possible combinations of pointers and by consequence, of possible MDS sequences. The average A/T-percentage of IESs gives a remarkable clustering principle for identifying the real pointer occurrences. It is well-known that the coding sequences are in general richer in C/G than the non-coding ones. We do not suggest that ciliates would recognize the A/T-percentage of IESs in *all possible combinations of pointers*. Rather, we point out the unexpectedly clear separation between the real gene and the vast majority of the other combinations, from the point of view of A/T concentration.

It turns out that when varying pointers of at least five nucleotides, the A/T separation is most clear, with the real assembly being one of very few with a high average A/T concentration per IES, in general over 80%. For non-scrambled genes, even pointers with four nucleotides give a clear separation. For shorter pointers however, this approach yields a larger cluster for the real gene, albeit it still discriminates against an impressively high number of alternatives. The result was especially unexpected in the case of pointer *aa* of the *R*1 gene in S.nova and in the case of pointer *ta* of Aminoacyl-tRNA synthetase cofactor on E.octocarinatus.

We observed that for highly scrambled genes like DNA polymerase alpha gene in P.weissei there are only a few number of successful assemblies. This suggests that such genes are very stable with respect to the variation of pointer positions. We think that this will permit a clear separation even for small pointers, however we could not verify this claim due to combinatorial problems.

Because of the large number of combinations in the case of short pointers, we could only investigate a subset of them. It is possible, as suggested by the case study on the alpha telomere binding protein in S.nova, that the cumulative effect of all combinations leads to clearer separation.

Fixing the position of short pointers to their real positions greatly reduces the complexity of the problem. This suggests that the template-based recombination mechanism is especially needed on pointers shorter than five nucleotides.

Acknowledgments The first and the third authors acknowledge the support of the Science and Technology Center in Ukraine, project 4032. The second author gratefully acknowledges support by Academy of Finland, project 108421.

References

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

[2] Prescott, D. M., Ehrenfeucht, A., and Rozenberg, G., Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates. *Journal of Theoretical Biology* **222** (2003) 323–330

[3] Angeleska, A., Jonoska, N., Saito, M., and Landweber, L.F., RNA-Template Guided DNA Assembly. *Journal of Theoretical Biology* **248**, Elsevier (2007), 706–720.

[4] Nowacki, M., Vijayan, V., Zhou, Y., Schotanus, K., Doak, T.G., Landweber, L.F., RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature* **451**, doi:10.1038/nature06452, (2008) 153–158.

[5] Cavalcanti, A., Clarke, T.H., Landweber, L., MDS_IES_DB: a database of macronuclear and micronuclear genes in spirotrichous ciliates. *Nucleic Acids Research* **33** (2005) 396–398.

Lemminkäisenkatu 14 A, 20520 Turku, Finland | 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 Sciences

ISBN 978-952-12-2109-5
ISSN 1239-1891