Nucleic Acids Research, 1995, Vol. 23, No. 11 2006–2013

Consensus inverted terminal repeat sequence of *Paramecium* IESs: resemblance to termini of Tc1-related and *Euplotes* Tec transposons

Lawrence A. Klobutcher* and Glenn Herrick[1]

Department of Biochemistry, University of Connecticut Health Center, Farmington, CT 06030, USA and [1]Department of Cellular, Viral and Molecular Biology, University of Utah School of Medicine, Salt Lake City, UT 84132, USA

Received January 19, 1995; Revised and Accepted April 26, 1995

ABSTRACT

During the formation of a transcriptionally active macronucleus, ciliated protozoa excise large numbers of interstitial segments of DNA (internal eliminated sequences; IESs) from their chromosomes. In this study we analyze the published sequences of 20 IESs that interrupt surface protein genes of *Paramecium* and identify a consensus inverted terminal repeat. This sequence is similar to the ends of the Tc1-related transposons found in nematodes and other metazoans, as well as to both the ends of the Tec transposons and at least some of the IESs in the distantly related ciliate *Euplotes crassus*. The results of these analyses bolster previous proposals that IESs were created by transposition.

INTRODUCTION

Ciliated protozoa undergo extensive reorganization of their genomes during their normal life cycles (reviewed in 1–5). This global DNA rearrangement process is possible because each organism contains both a micronucleus and a macronucleus. The micronuclear genome consists of conventional eukaryotic chromosomes. It is transcriptionally inactive during the vegetative proliferation of the organism, but plays a major role during sexual reproduction and thus is viewed as a 'germline' nucleus. The second nucleus in the cell, the macronucleus, provides for the transcriptional needs of the cell during vegetative proliferation. Its genome consists of a subset of the DNA in the micronucleus, arranged as subchromosomal, linear DNA molecules. These are each present in 45–15 000 copies per macronucleus depending on the species. Following the sexual phase of the life cycle, the old macronucleus is lost and a new one is generated from a mitotic copy of the micronucleus. Extensive DNA rearrangement occurs during the development of the new macronucleus, including chromosome fragmentation, *de novo* telomere synthesis and DNA amplification.

In addition, macronuclear development entails the elimination of large numbers of interstitial segments of DNA (IESs; internal eliminated sequences) through DNA breakage and rejoining, or splicing events. Such events are extremely common in hypotrichous ciliates, where 50 000–100 000 IESs are excised during macronuclear development. All ciliate species examined in detail have been found to undergo DNA breakage and rejoining events, but characterization of different IESs and their excision processes have generally indicated that there is a great deal of heterogeneity within and among species. For example, hypotrichous ciliate such as *Oxytricha nova*, *Oxytricha fallax*, *Oxytricha trifallax* and *Stylonychia lemnae* all possess relatively short IESs (14–550 bp: e.g., 6–10; A. Seegmiller, K. R. Williams, R. L. Hammersmith, T. G. Doak, D. Witherspoon, T. Messick, L. L. Storjohann and G. Herrick, manuscript submitted). These IESs are all bounded by short direct repeat sequences, one copy of which is retained in the macronuclear DNA, but the length and sequence of the direct repeats varies from IES to IES. IESs in *Tetrahymena thermophila* are generally larger, but again differ in the lengths and sequences of their direct repeats (e.g., 11–14). Moreover, for one *T.thermophila* IES, a purine-rich sequence element flanking an IES has been shown to be essential for developmental excision, but other characterized *T.thermophila* IESs lack this element (15,16). Large transposons have also been identified that undergo developmental excision. In the hypotrich *Euplotes crassus*, two related 5.5 kb transposons (Tec1 and Tec2) exist in ~7000 copies each per haploid micronuclear genome and undergo developmental excision (17–19). Approximately 2000 copies of a related transposon (TBE1) exist in the micronuclear genomes of *O.fallax* and *O.trifallax*, but these differ in that they have telomeric repeat sequences at their ends and also appear to be excised by a different process than the *E.crassus* Tec elements (20,21). Finally, although most IESs are excised such that the immediately flanking DNA sequences are joined, some macronuclear DNA molecules in *O.nova* appear to be formed by a splicing and deletion process that reorders micronuclear DNA segments (e.g., 22).

E.crassus is one ciliate system for which a somewhat unified picture of developmental DNA breakage and rejoining has emerged. This organism possesses both short, unique sequence IESs (17,23) and the Tec elements mentioned above. All of these sequences are bounded by the same dinucleotide repeat, 5′-TA-3′,

* To whom correspondence should be addressed

and are precisely excised during development (24). Moreover, concomitant with the excision of both Tec elements and IESs in *E.crassus*, free circular forms appear with an unusual junction region, indicating that the two classes of sequences are excised by a similar mechanism (25,26). These observations have bolstered the proposition that the short IES sequences represent degenerate forms of transposable elements subject to developmental DNA excision (reviewed in 3,4). It has been noted that the micronuclear genome presents a favorable site for the expansion of a transposon family because the micronucleus is primarily transcriptionally inactive. Thus, any transposon that developed a mechanism for excision (with either element or host encoded functions) during macronuclear development would be able to extensively populate the micronuclear genome, and, over the course of evolution, some copies of the transposon would be expected to degenerate to the point where they retained only the essential *cis*-acting sequences required for excision.

Recently, studies of micronuclear gene organization in *Paramecium* have provided one of the first indications of inter-specific similarity in the IES phenomenon. Numerous short IESs (28–882 bp) have been identified by a number of different laboratories in members of the multi-gene family encoding a major surface protein of this species (i-antigens), and all have 5′-TA-3′ terminal direct repeats as in *E.crassus* (27–29; H. Schmidt, personal communication). In the current study we have analyzed the terminal sequences of these *Paramecium* IESs. Our results indicate that the *Paramecium* IESs have a region of conserved sequence at their ends, in inverted orientation, and that this sequence is similar to the ends of the Tec transposons and the Tc1-related transposons. The deduced consensus sequence for the *Paramecium* IESs is shown to have predictive value for identifying IES ends. Additional analyses suggest that *Paramecium* IESs have lost internal sequences during evolution, such that they approach a minimum size of 28–29 bp. These results provide additional evidence for the hypothesis that the small IESs arose via the duplication of transposable elements.

MATERIALS AND METHODS

DNA sequence analysis

To search for significant sequence similarities within and flanking IESs, we employed a previously described statistical method (30,31) that compensates for the overall base composition of the region under analysis. In brief, one calculates the overall base composition of a series of sequences aligned relative to a reference point. The average expected occurrence of a given base (M) at a particular position is then calculated as Nf_b, where N is the number of individual sequences being analyzed and f_b is the overall frequency of an individual base (A, C, G or T), or two alternative bases (A or G, C or T, A or T, etc...). The standard deviation is calculated as the square root of $Nf_b(1-f_b)$. Positions in the sequences where a base, or two alternative bases, occurred at >M + 3 standard deviations were considered to be statistically significant.

Consensus sequence

To derive a consensus sequence, only those positions were considered where a base, or a combination of two alternative bases, was present at a statistically significant frequency as defined above. In some cases, a single position displayed two or more statistically significant bases and/or combinations of alternative bases. In this instance the following conventions were used for deriving the consensus: (i) the statistically significant single base was included in the consensus when it was present at a particular position in ≥70% of the sample sequences, and (ii) in cases where a single base was not present in 70% of the sample sequences, the statistically significant combination of two bases that occurred at the highest frequency was included in the consensus.

IES profile analyses

The programs PROFILEMAKE and PROFILEGAP (GCG, version 8.0; ref. 32) were used to construct position-specific scoring tables, or profiles, from aligned *Paramecium* IES ends, and to search micronuclear sequences for profile matches. PROFILEMAKE settings were default, except GG and CC match scores were elevated from 1.0 to 1.5; other scores were default (AA and TT matches 1.0, mismatches –0.6). Profile B was derived from the 15 terminal nucleotides of each of the two ends (ungapped) of the *Paramecium* IESs displayed in Figure 1. This profile was used to search for IESs in non-*Paramecium* micronuclear DNA sequences. Profile A was constructed from the same regions of the IESs, but IESs –10, 1416 and 6435 of the 51A gene and IESs –11, 1417 and 5464 of the 51 B gene were omitted. Profile A was used to search *Paramecium* micronuclear DNA sequences. For both profiles, the constituent sequences of identified paralogous IES pairs were given reduced weights, W, reflecting the fraction, f_m, of their first 15 nt that match. The weighting factor was calculated as $W = 1 - 0.5f_m$. The paralogous IES pairs that were weighted are the three shown in Figure 3 (included in profile B only) plus the 28 bp α-51D and 28 bp ε-51D IESs (see Fig. 1A).

PROFILEGAP was then used to search micronuclear sequence files for matches with the profiles. Default settings were used except global matches were demanded, where the highest-scoring full 15 bp match is reported. Because PROFILEGAP only reports the top match in a given sequence, the following procedure was employed to examine successive next-best matches. The first 5 nt of the top matching sequence (which were always TA and three further nucleotides) were altered to 'NNNNN', rendering that sequence a poor match with the profile. The resulting 'damaged' sequence was then searched for the next best match, and in turn the first 5 nt of this match were altered. In this fashion the highest-scoring 0.3% of all matches were identified [a sequence of N nucleotides has $N - 14$ possible independent matches with a 15 nt profile, so 0.3% of all possible matches = $0.003 \times [N - 14]$ matches]. To be able to search for both ends of each IES, each profile query sequence was generated by appending the reverse complement of each IES-carrying micronuclear sequence 3′ to the sequence itself, so the full query sequence was a total palindrome, twice the length of the micronuclear sequence. No gaps were found in any match.

RESULTS

Sequence similarity at the ends of *Paramecium* IESs

Recently it has become possible to analyze the micronuclear organization of genes in *Paramecium*. Five genes encoding variable surface proteins have been the first to be sequenced, and 20 IESs have been identified within their coding and flanking

A.

```
Gene: IES          -60       -50       -40       -30       -20       -10        1        10        20        30        40        50
51B:left -11(36 bp)   AGTTTTGTAAAAATTAAATTTTATTCATAATTTAGCTTGATTAGAATGTAAACCTAAAGTTATAGTATAATCCCATCAT
51B:right -11(36 bp)  ACTTGCTGCTAAAGTTAGCAACAATGCAAATATCATGAATTATTTATTCATTAAAAGTATTATAGTTGAAATTAGAGAT
51B:left 1417(44 bp)  TTTCTGGGATTCAGCCAGTTCAACTTGCAAGGATAAAACTTGTGCTAATGCTCCATCAACTATAGTAATTTTATAATAATATT
51B:right 1417(44 bp) CATTGGTTGAGTTTACTGTACAAGATGATAAAAAGGCAACACATAAATCATTAGTAGTATTATACCTATATAATTATTATATA
51B:left 3928(416 bp) AACTGGTTGCGTAACAATTTAATCTGCATGCACTGGATACACAACAATTGCTAATTGTTATACAGCTTAATTAAGAAACATTTTCAACAATTTTGATATTAATCAAAAGG
51B:right 3928(416 bp) CTGATATAGCTTAACAGGCAGAATCACCACTATTAGCGGTACAATATCCAGCAGTTGATCTACAGCTGATTAATTTAATTTATAAGTGTATAATTTTAAATTTGTTAATA
51B:left 5464(28 bp)  TTCATCAGCAAATGTATAATGTGAATGGATGCCAGCTGTAGGAACTAACTAAGCATATTGTATTAATAAATTTAT
51B:right 5464(28 bp) AACTTGCACAAGCTGATTCAGAGGTTAATGTTGCTGGGGCTGTATTACAAGTCTTAACAGTATGGAAGAAATAAA
51A:left -712(77 bp)  AGACATGTATCAAAATGCAGGCAATTTCCAATATATTTTTCGAGAATATTCTATTGTTAGTATTGCTACTTTAAAAATCTTCATCTTCATGAATTAATA
51A:right -712(77 bp) CAAACTTTCTTGGTTTCCAAACAAGAAATTTTCCATAAAATTCCATTGATTTATTATTAATATTATCGAGTTGTTGAATAAATATATAGATATTTTTTT
51A:left -10(29 bp)   AAAAAAAAAAAAAATTAACAAACTAACTATTTAAAAATTCATTCGATTCAAATTTAATTTTATAGTTAAGTAATC
51A:right -10(29 bp)  GGCTTGCCGCTAATGCAAGCATTAATGAAAGTATGAAGAATTTCTAATTCATTAAAGTATTATAGTTAAGGACAT
51A:left 1416(74 bp)  CTTCTGGGATTCTGCTAGTTCAAGTTGCAAAGATAAAACTTGTTTAAATGCTCCATCAACTACAGTTATTATTTAATATTTTCCAAATTAATTATCCA
51A:right 1416(74 bp) CACTTGTTGAATTAACTGTGCAAGTATTTAAGAATGCTTAACAAAGGTCGTGTGTGGTGTTATAGTCGAAGAGAGAGATTTAATATAGAAAATTATTT
51A:left 1835(28 bp)  ATAAGGCTTGCTCTACAGCCTCAAAGACCTTTTCAACTACTACATAATGCTAAACTCATTTATAGATGGATTGTT
51A:right 1835(28 bp) ACATCCTTGAATAGTTACTGATCCATTTACAGTTGCAGGGTTATTAGCAACACATGATGATATAGATACTTGGAA
51A:left 2591(370 bp) ATTACTAAACACCAAGCGAAACATGCACAGTCGTTTATAAAGTTGGAGCACTGGGATGTGTATTCTAAAATTAACAAAAATACATTATTTTACTTTTTATAAATATTTGC
51A:right 2591(370 bp) TGTGAGTGTTTTATGGCATTAAGCTTGTGTCATATAGTCTGTACAATTAGCTGATCGTGGTATTGAGAAGTTAAGAAATAAAATGATGGATTATTCAATAATTTTAATCG
51A:left 4404(77 bp)  AGGTTTGAATGATAGTAAATGTTCAGCTTACAACGCAGCTTGCACATCTCTAGTTGATGGTATTGTGGTGTTATTTTAAAAAGTAATATATTTTTTAAA
51A:right 4404(77 bp) AAATACATTTATTTTGTGTGGTGTAATCTTTACAGTTTGCTTTTTCCTCTTAACAAGCAGTATGGTAGGTTTAAGTTTTTTTTTATGAGAAAACCAAATA
51A:left 4578(882 bp) TGCAGTAAGTTGCAGTTCAATAACTGGAACTTCTCTTGATCATACAAAATGCCAAGCCTATACAGCTAACGTAAGAAATTTTATTGTAAATATATTTTCAGCTAGATTTA
51A:right 4578(882 bp) CACATGTAGTCTTAAAATCTTAGCATGTTGTACCATCTGTAATTGAAGTACATTTAGTGTTACAGTCTAGGAAATGAAATTTTTTGAATTGAATTAATTCACAAATTAAG
51A:left 6435(28 bp)  ATGTGAATGGATACCAGCTGTTTTGAATGCATAAAATGTTGTTACTTCCCCTGCATATTGTATCGATAATATTGT
51A:right 6435(28 bp) AACTAGCACAAGCTGCTTCGCTTGTTAAAGTTAGAGGTGCAGTACTACAAGTCTTTATTGTATAATATATTAATA
51A:left 6649(370 bp)  TTAAATGGTACTGTTTGTGCTTGGGATAGTGCATAAAATAAATGCAGAGATAAAGATTGCTACAGCTGAGCATAAAAAATTTTAATTTATAATATAGGAAATGCCCTGTT
51A:right 6649(370 bp) CCAGCAGTACATCCAGCTCTCTAAGTTTAGCATGCTGCATGTGTAGTTCCACTGAAATCCTACAGGTTAAAAAAAATTTTTTTCTATTTTCTTTTAGTTCAGACTTTAAG
51A:left 15885(54 bp)  TCTTATATTGGATGATGATTAAAATATTATATAAGTAGCGGCGTCTACTGAAACATATTCTACAGTTTAGAATTCTCACCTGATAAAA
51A:right 15885(54 bp) GCATTCAAATATCGTTTTGAATAAAGAAGCAACAAAATAATCATTGAAGGACAAGCCGTTTACAGCATCTTGAGTTGTAGCNNNAAAA
51A:left 15637(47 bp)                                                                    TATAGTCCTAATTGTGTAATTTNA
51A:right 15637(47 bp) ATAATCTTCAAAACTTTAAGAGATTTGAATTATGGTGTGCATGTTCAAACGTGTTTGAAGTAAAGCATGTCAGAACAAAATTAG
156G:left 76 bp        GTTTTCACTCATTTTAATTCAGCAATTATACTCTTAACTTTTTAAATAAAAAAAATTAGTTACAGCAATATAAAATATATTTTACATTCTCTAATTTCA
156G:right 76 bp                                                                         TATGGTAATAATAATTCTCTGTTTTTTTTATAATTTCAA
156G:left 67 bp         TGATTCTTTAAATAACTCATCAAAAACTAATAGTTTCTTATTAATTAGATTAAGTCTCAATACTGTTGATTAAGGTAATTTAATAAATTTTCTA
156G:right 67 bp        ATGAAATCTGAATTAAAACTTCGGACGAGTTTGATTGCTAGAAACTTATTTGGATAAGATTAGAGTAGGTCAAAAAAAAAAATTATATAAATTT
α-51D:left 28 bp        AAGCCAGAATTAAAACATAAGATGGAGTAGATGATTCAGGAAATCCCAAATACATTTATGTATACTAAAAGTAAT
α-51D:right 28 bp       GGTTAAATCTGAACAAGCCAAATCTTTGCAAGCATTATTCAACCAACCACATTTCCTGTTATAATTAATAGATA
α-51D:left 44 bp        AAGCAGGAACTTGTGAAAATATTCCAAGCCCAACCAGTTAAGCTGATTGTGATATACATTTATAGTAATTTGAAAAATCATCT
α-51D:right 44 bp       CAGCAGCAAGAGCAGTAACTTTTGGTGTTCTACATTTACTAGCTGAGAACACACACCAATTATAGTTATTGTAAATATATTTT
ε-51D:left 28 bp        AAGCAAGAATTAAAACATTCGATAAAAATGATGAAAACGGAAATCCATTATACATTTATGTATGCTAAAAGTAAT
ε-51D:right 28 bp       GGTTAAATCTGAACAAGCCAAATCTTTGCAAGCATTATTGAACCAACCACATTTTCCTGTTATAATTAATTATTA
```

B.

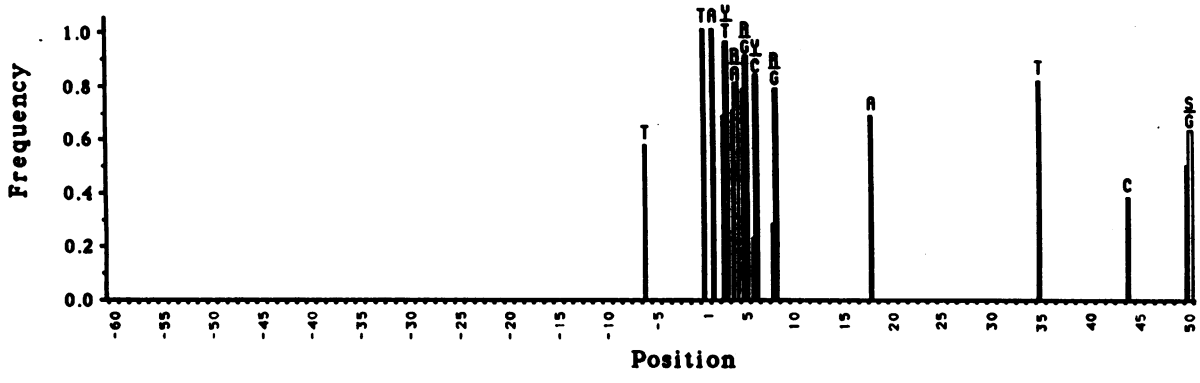

Figure 1. Analysis of *Paramecium* IESs and flanking regions. (**A**) The ends and flanking regions of the 20 *Paramecium* IESs included in the analysis are shown. The sequences are aligned at the 5′-TA-3′ direct repeat (bold type) that flanks each IES (note that one of the repeats is technically not part of the IES, as it is retained in the macronucleus following excision). The 60 bp flanking each IES are negatively numbered, while up to 50 bp of the IES ends are positively numbered. For the *P.tetraurelia* 51A (27; GenBank accession no. L26124) and 51B (29; GenBank accession no. U07603) surface protein gene IESs, the designations of the IESs correspond to those used in the original description of the IESs, and refer to the position of the IES in the macronuclear DNA sequence. The sizes of these IESs are shown in parentheses. For the *P.primaurelia* 156G (28) and *P.tetraurelia* 51D (Schmidt, unpublished results) surface protein genes, the IESs are designated by their sizes. Left and right ends of the IESs are based on their orientations in the previous publications. Complete flanking sequences were not available for the left end of the 51A gene IES 15637 and the right flanking region of the 156G gene 76 bp IES. (**B**) A histogram displaying the results of the statistical analysis of base composition. Base frequency versus positions within the IESs and flanking regions is displayed. Only those positions where a statistically significant deviation from random base composition (i.e., > expected mean + 3 S.D.) are plotted. Significant positions for A, G, C and T are shown as solid black bars, while significant positions for a combination of two bases are shown as open bars. For simplicity, in cases where a single position displayed a significant deviation from random base composition for more than two bases and/or combinations of bases, only the base and two base combination with the highest frequencies are shown. The overall base frequencies for the flanking regions are: G = 0.17, A = 0.35, T = 0.32 and C = 0.16. For the flanking regions, where $N = 38$, a base or combination of two bases had to be present at a given position in excess of the following values to be considered statistically significant: G > 13, A > 22, T > 20, C > 12, G or A (R) > 29, C or T (Y) > 27, A or T (W) > 34, G or C (S) > 21, G or T (K) > 27, and C or A (M) > 28. The overall base frequencies for the IES regions are: G = 0.11, A = 0.40, T = 0.41 and C = 0.08. For the first 15 positions of the IESs, where $N = 40$, a base or combination of two bases had to be present at a given position in excess of the following values to be considered statistically significant: G > 10, A > 25, T > 25, C > 8, R > 29, Y > 29, W > 39, S > 15, K > 30 and M > 28.

Figure 2. Comparison of the deduced consensus sequence for the *Paramecium* IESs with the sequences at the ends of the *E.crassus* Tec1 and Tec2 transposons (26) and with the consensus for the Tc1-related transposons (34–37). Identical bases are highlighted with a black background, and similar positions (e.g., G versus R) are highlighted with a stippled background. R = G or A, Y = C or T, K = G or T and S = C or G.

```
51A  -10     caaatttaatttTATAGTTAAGTAATCA.......ATGTCCTTAACTATAatacttt.aatg
51B  -11     taaacctaaagtTATAGTATAATCCCATCATATCTCTAATTTCAACTATAatacttttaatg

51A  1416    ccatcaacTACAGTTATTATTTAATATTTTCCAAATTAATTATCCAAAATAATTTTCTATATTAAATCTCTCTCTTCGACTATAacaccacacacg
51B  1417    ccatcaacTATAGTAATTTTATAATAATAT.....................TTATATAATAATTATATAGGTATAatactactaatg

51A  6435    cctgatattgTATCGATAATATTGTTATTAATATATTATAcaataaagacttg
51B  5464    aagcatattgTATTAATAAATTTATTTTATTTCTTCCATActgttaagacttg
```

Figure 3. Pairs of IESs interrupting paralogous surface protein genes 51A and 51B of *P.tetraurelia*. Sequences were aligned to exclude gaps from ends. Vertical bars connect matching nucleotides. IES sequences and flanking TA direct repeats are in upper case letters; macronuclear sequences are in lower case (one TA repeat is retained in the macronucleus). The underlined sequences at the top of the first alignment represent the start codons for the 51A and 51B genes.

regions (Fig. 1A; 27–29; H. Schmidt, unpublished results). The variable surface proteins represent a group of related major cell membrane proteins that are generally expressed in a mutually exclusive manner under different environmental conditions (33). The genes encoding these surface proteins constitute a multi-gene family that appears to have arisen by duplication of an ancestral gene.

To determine if significant sequence similarities exist within or around *Paramecium* IESs, we aligned each end of the 20 *Paramecium* IESs relative to the 5′-TA-3′ direct repeat that flanks each IES (Fig. 1A), and then performed a statistical analysis on each position to look for non-random base composition. In the alignment, we included the 60 bp flanking each IES plus up to 50 bp of the IES itself (in cases where the IES was <100 bp long, the IES was divided into two equal halves, omitting the central base for IESs with odd numbers of bp). Nucleotides found at a particular position more often than predicted by the local base composition were then identified by a statistical method (see Materials and Methods). Furthermore, separate analyses were carried out on the flanking regions and on the IESs *per se*, as the flanking regions usually represent coding regions that are more GC-rich (33% GC) than the IESs (19% GC; note that this is a conservative approach that minimizes the possibility of particular positions appearing to have non-random base composition simply because of an unusual overall base composition of a sub-region). In these analyses, the expected mean occurrences for each base (G, A, T or C), and for all possible combinations of two alternative bases (G or A, C or T, A or T, etc...), were calculated along with standard deviations. Particular positions where the occurrence of a given base, or two alternative bases, exceeded the expected mean occurrence by three standard deviations were then considered to have statistically significant deviations from random base composition.

This analysis identified a cluster of such positions at the ends of the *Paramecium* IESs (Fig. 1B). This region includes the first 8 bp of the IES ends, beginning with the 5′-TA-3′ direct repeat. A few additional significant positions are present in the IESs and flanking regions, but these are isolated and, hence, their biological significance is dubious. Based on this analysis, a consensus sequence of $5'\text{-}T_{1.00}A_{1.00}Y_{0.95}A_{0.70}G_{0.78}Y_{0.83}NR_{0.78}\text{-}3'$ (subscripts denote the frequencies of the bases in the IES sample) was deduced for the *Paramecium* IES ends. This result has a number of implications. First, it implies that *Paramecium* IESs have inverted terminal repeats, although the repeats for an individual IES are often imperfect. Secondly, it indicates that the IESs are at least functionally related, and have perhaps derived from a common ancestor. Finally, as shown in Figure 2, the deduced consensus sequence for the *Paramecium* IESs is very similar to the ends of the *E.crassus* Tec elements (26), as well as to the ends of the family of Tc1-related transposable elements (34–37), suggesting that the *Paramecium* IESs are also related to these transposons.

We also note that the sequence similarity at IES ends may extend internal to the first 8 bp. The statistical analysis was also repeated using a significance cut-off of the expected mean plus two standard deviations (not shown). Many more statistically significant positions were detected throughout the regions analyzed, but a cluster of significant positions was still discernible at the IES ends, and in this case the cluster extended to position 14.

Conserved features of IESs in paralogous surface protein genes

An additional indication that the termini of *Paramecium* IESs are important was obtained from an analysis of IESs in paralogous surface protein genes. Scott *et al.* (29) have noted that three pairs of IESs (Fig. 3) exist at the same locations in the micronuclear

copies of the 51A and 51B surface protein genes. This implies that these pairs of IESs were present in an ancestral gene prior to the duplication event that gave rise to the present day loci. As such, the two IESs of a pair diverged from a common ancestor and also represent paralogous sequences. Sequences within such paralog pairs that are required for function are expected to be conserved, while non-functional sequences should be subject to both mutation and deletion.

Scott *et al.* (29) have noted a trend towards terminal sequence conservation for paralogous pairs of IESs and we present a more thorough analysis in Figure 3. The sequences of the paralogous IESs have generally diverged, but there is a clear trend towards conservation of terminal sequences. This is most evident for both ends of the 51A –10/51B –11 IES pair, the left ends of the 51A 1416/51B 1417 IES pair, and the left ends of the 51A 6435/51B 5464 IES pair (Fig. 3). In each of these cases, at least seven of the nine terminal bases are identical. The remaining two IES ends show conservation only of the terminal 3 or 4 bp. Overall, the results indicate that it is only the terminal sequences that have been evolutionarily conserved.

It is also noteworthy that the members of two of the paralogous IES pairs have undergone changes in size. The 51B –11 IES is 36 bp long compared with the 29 bp 51A –10 IES, while the 51A 1416 IES is 74 bp long compared with the 43 bp 51B 1417 IES. This suggests that the IESs have undergone various degrees of loss of internal sequences during evolution (it is also formally possible that one member of the paralogous pair has gained sequences during evolution), and may be converging on the minimum IES size. There is also some indication of a minimum size for IESs if one considers the entire size distribution of the 20 IESs shown in Figure 1A. The IES lengths are 28, 28, 28, 28, 28, 29, 36, 44, 44, 47, 54, 67, 74, 76, 77, 77, 370, 370, 416 and 882 bp. It is thus evident that there is a cluster of IESs in the size range of 28–29 bp. This suggests that smaller IESs (i.e. <28 bp) cannot be excised efficiently either because of steric constraints or because they do not possess sufficient sequence information to specify excision.

The predictive value of profiles of *Paramecium* IES ends

The computer programs PROFILEMAKE and PROFILEGAP (32) were also employed to determine if the sequence similarity at the ends of the *Paramecium* IESs was significant and if sufficient sequence information was present to detect IESs in the micronuclear DNA of *Paramecium* and other ciliates. PROFILEMAKE evaluates a series of aligned DNA sequences to generate a position-specific scoring table, or profile. A profile is, in effect, a weighted consensus in which a nucleotide that is frequently present at a position receives a high value while an infrequent nucleotide receives a low value. PROFILEGAP then uses this profile to search a query sequence (in this case a micronuclear DNA sequence) for the highest scoring match. In the current analyses, default program parameters were used in constructing profiles and conducting the searches except that: (i) GC base pairs were given an increased weighting because of their scarcity in IESs, and (ii) paralogous IESs received a reduced weighting in constructing profiles to avoid over representing the information present in these closely related sequences (see Materials and Methods for details).

Initially we sought to determine if the information present at IES ends was efficient in identifying IESs within *Paramecium* micronuclear DNA. For this purpose a profile (profile A) was constructed from the first 15 bp of all of the IES ends shown in Figure 1A, except the ends of the three pairs of paralogous IESs in the 51A and 51B genes. Profile A was then used to search the complete micronuclear 51A and 51B gene sequences for the omitted IESs. This approach was taken because sequences used to construct the profile can 'vote for themselves' (38,39), and this strategy avoids this potential problem. The top 0.3% matches to the profile were determined for the complete sequences of both the micronuclear 51A and 51B genes (Table 1). A total of 12 ends of the paralogous IESs are present in the A and B genes (six in each), and eight of these were detected among the top 0.3% matches. If one considers all IES ends present in the 51A and 51B genes, including those used to construct the profile, 20 of 26 were detected in the top 0.3% matches (Table 1). Thus, the profile is effective for identifying *Paramecium* IES ends. It is not, however, a perfect predictor, as all IES ends were not identified and a significant number of sequences that are not destined for development excision were detected by the profile (e.g. of the top 0.1% matches in the 51B gene, five were bona fide IES ends and 12 represented non-IES ends).

We next asked whether a profile derived from *Paramecium* IES ends was effective in detecting IES ends present in the micronuclear DNA of another organism. For this purpose a second profile was constructed (profile B) from the first 15 bp of all of the *Paramecium* IES ends, and used to search the micronuclear LEMICV sequence of the hypotrichous ciliate *E.crassus* (17,23; GenBank accession no. M28500). This micronuclear DNA sequence contains five small IESs, all of which are bordered by 5′-TA-3′ direct repeats. The profile detected two of the 10 IES ends, both within the top 0.1% of the matches (Table 1). Although the success rate was depressed in this case relative to *Paramecium*, it should be noted that two of the IESs in the *E.crassus* LEMICV sequence appear to be unusual, and perhaps a distinct class, as they display an unusual base composition and are excised later than the other IESs (23). If these two IESs are discounted, two of six possible *E.crassus* IES ends were detected.

Profile B was also used to search a sequence file representing the complete *E.crassus* Tec1 and Tec2 transposable elements (40; GenBank accession nos L03359 and L03360). The two Tec2 ends, which are identical, were detected among the top 0.1% of the matches, but the Tec1 ends (also identical to each other) failed to appear among the top 0.3% of the matches (Table 1). Detection of the Tec2 ends but not the Tec1 ends was initially puzzling, as the first 9 bp of the two elements are both identical and universally conserved (26). Further examination of the profile and sequences indicated that this result is likely due to the fact that *Paramecium* IESs and Tec2 elements usually have A residues at positions 13 and 14 (Fig. 1A; 26), while Tec1 elements frequently have T residues at these positions. Thus, there is some indication that the *Paramecium* IESs share a closer relationship with Tec2 as compared with Tec1.

Finally, as a negative control, the *O.nova* micronuclear sequence of the R1 gene (7) was searched with profile B. This sequence contains five IESs, but none of them is bordered by 5′-TA-3′ direct repeats. None of these IES ends was detected in the top 0.3% matches (Table 1). Thus, the profile derived from the *Paramecium* IESs is not effective in detecting all IES ends, but displays specificity toward developmentally eliminated DNA sequences that are bordered by 5′-TA-3′ direct repeats.

Table 1. Results of IES profile analyses

Micronuclear sequence searched	Length (nt)[a]	Total IES ends present	IES ends detected among top profile matches			
			0–0.1%	0.1–0.2%	0.2–0.3%	Total (0–0.3%)
Paramecium tetraurelia 51A gene[b]	22 270	6 (18)	2 (9)	2 (4)	0 (1)	4 (14)
Paramecium tetraurelia 51B gene[b]	17 214	6 (8)	3 (5)	0 (0)	1 (1)	4 (6)
Euplotes crassus LEMICV locus[c]	7570	10	2	0	0	2
Euplotes crassus Tec1 & Tec2 transposons[c]	21 262	4	2	0	0	2
Oxytricha nova R1 locus[c]	4070	10	0	0	0	0

[a]The stated lengths of sequences represent both DNA strands.
[b]Profile A was used to search for IES ends. Values given are for the paralogous IES ends only, while values in parentheses include all of the IES ends.
[c]Profile B was used to search for IES ends.

DISCUSSION

Relationship of *Paramecium* IESs to IESs in other ciliates and transposons

Both the statistical and profile analyses lead us to conclude that the *Paramecium* IES ends consist of similar inverted terminal repeats, with a consensus of 5′-TAYAGYNR...3′ (Fig. 1; Table 1). E. Meyer (personal communication) has independently reached a similar conclusion using a somewhat different statistical approach. We believe that at least some of the nucleotides represented in this consensus sequence interact critically with components of an IES excision system (see below). How did these 40 IES ends evolve to have this similarity? There are two obvious possibilities, in a clear parallel to the on-going debate over the early versus late origin of spliceosomal introns (see 41 and references therein). On one hand, these IESs may have evolved from sequences that existed in their respective genomic locations since the initial creation of the genes they interrupt. In concert with the elaboration of a developmental DNA excision system, sequences whose removal was beneficial to the organism would have come under selection to retain adaptive mutations at their ends that enhanced excisability. That is, the IES ends are similar today because they converged from non-homologous sequences. Alternatively, these IESs may all be homologous (paralogous) because they arose from a single IES sequence through a series of duplications, most plausibly by transposition. Thereafter, these sequences suffered mutations, and purifying selection eliminated IES end mutations that hampered excision. That is, the IES ends are similar today because they are homologous, and the similarity reflects conserved features retained from the original IES sequence.

It is difficult to distinguish between the adaptive convergence and constrained divergence explanations. However, several aspects of our results support the duplication–divergence–conservation alternative. First, of the 20 *Paramecium* IESs we analyzed, at least some are clearly homologous because they interrupt identical positions in paralogous genes (Fig. 3). In each of these three cases we feel safe in concluding that the two members of a pair have diverged from a common IES sequence and that the end sequences have been conserved (while central sequences have diverged and been deleted down to a minimum length of 28–29 bp). Secondly, the *Paramecium* IES consensus sequence bears strong similarities to the ends of the *E.crassus* Tec transposons and Tc1-related transposons (Fig. 2; Table 1). Like all transposons, these elements form paralog families in their host genomes. The Tec and Tc1-family elements have previously been related by virtue of possessing homologous transposase genes (39), and form part of a large transposon family that includes *Oxytricha* TBE1, a subfamily of IS*630*-related bacterial elements, and the widely distributed eukaryotic *mariner* subfamily. Like the *Paramecium* IESs, most of these transposons are bounded by 5′-TA-3′ direct repeats that are likely the result of target site duplications upon insertion (39,42). Furthermore, several of the eukaryotic elements in this large family have a propensity for somatic excision (39). The ciliate Tec and TBE1 elements undergo an extreme form of somatic excision, as all copies are excised during macronuclear development (17–20). The similarity of the *Paramecium* IES ends to those of some members of this family thus suggests that the *Paramecium* IESs were generated by an element of this family, and that the *Paramecium* IESs are homologs.

These results provide additional support for the previously suggested hypothesis (reviewed in 3,4) that IESs are ancient copies of transposons that were both transpositionally active in the micronuclear genome and also able to undergo controlled excision during macronuclear development. One scenario is that, over the course of evolution, some copies of the transposon lost sequences not required for developmental excision to give rise to the current IESs. It will clearly be of interest to determine if *Paramecium* harbors a large transposable element that is related to the smaller IESs.

Our study, indicating that the *Paramecium* IESs are related to both the Tec elements and at least some of the IESs of *E.crassus*, represents one of the few instances where clear similarities in sequences subject to development elimination have been noted for different species of ciliates (also see 10,43). We also note that a recent analysis of multiple alleles of an IES with 5′-TA-3′ direct repeats in *O.fallax* and *O.trifallax* indicates that its terminal sequences are conserved, and that these terminal sequences are quite similar to the *Paramecium* IES consensus (A. Seegmiller, K. R. Williams, R. L. Hammersmith, T. G. Doak, D. Witherspoon, T. Messick, L. L. Storjohann and G. Herrick, submitted for

publication). Similar analyses of the three other IESs identified in these species, which have direct repeats of different length and sequence, also indicate that only terminal sequences are conserved, but the conserved sequences differ among these IESs and are unrelated to the *Paramecium* consensus. Thus, these results suggest that only a subset of the IESs in *O.fallax* and *O.trifallax* are related to the 5′-TA-3′ direct repeat IESs of *Paramecium* and *E.crassus*.

There are at least two possible ways to explain these various inter-specific relationships based on the transposon-based scenario suggested above. First, the transposon precursor of these IESs may have already existed in the common ancestor of these various species. Alternatively, there may have been multiple independent invasions of the micronuclear genomes of these species by related transposons. The latter scenario is not unreasonable, as at least some members of the Tc1-*mariner*-IS*630* family, the *mariner* subfamily, appear to have independently invaded a wide range of species (44). The analysis of IESs in orthologous genes in different ciliate species could help resolve this issue.

Finally, we must emphasize that although the current data provide an indication of the relatedness of deleted DNA segments in different ciliates and their origin from transposons, it is still difficult to support a unified picture for the origin of all IESs. First, there are differing views as to whether the hypothesis that IESs are derived from transposons is reconcilable with the observation that flanking DNA sequences are necessary for one particular DNA excision event in *Tetrahymena* (2,3,16,21). Secondly, as discussed more thoroughly in the Introduction, there remain many IESs in ciliates that vary in their terminal sequences and, hence, appear to be unrelated to the group defined in this study. Thus, it is currently unclear whether these other IESs have arisen from independent transposable elements present in ciliates or if they arose by a completely independent mechanism.

Role of the conserved IES termini

The sequence conservation at the IES termini implies that they are involved in some required function. We suggest that these sequences may play a role in either the developmental excision process or in transposition, as terminal sequences frequently play a major role in other transposition and site-specific recombination systems (reviewed in 45). Of these two possible functions, a role in excision seems to be the more likely. Excision must occur to generate functional macronuclear genes, so that there is continual selection to maintain features of IESs required for excision. Transposition of IESs on the other hand is unproven at this point, and we can envision no selection that would maintain the ability of IESs to transpose, unless, as discussed previously (21), excision and transposition are related processes.

One obvious question that arises is whether the sequences present at the IES ends are sufficient to specify excision. At face value, the computer profile analyses seem to indicate that the information at IES termini is insufficient to uniquely identify IESs, as not all IES ends were detected among the top profile matches and many non-IES ends received high scores. This might represent a failure of the profile procedure to accurately weight the positive and negative contributions of particular positions within the terminal sequence. That is, the procedure may not accurately simulate the features recognized by the excision machinery. On the other hand, the IES terminal sequence may be necessary but, on its own, insufficient to specify excision. Additional *cis*-acting sequences would then be necessary for uniquely identifying IESs for excision. Such additional sequence information could, for instance, simply involve having a second IES end in the proper orientation and within a specified distance (we attempted to address this possibility using 'two-ended' profiles, but the results were difficult to interpret). Alternatively, an additional sequence element involved in recognition and excision may exist either internal or external to the IESs. Such sequence elements might be analogous to the enhancer elements that have been defined for other site-specific recombination systems (e.g. 46) or the flanking sequence element that has been shown to be necessary for the excision of an IES in *Tetrahymena* (15,16). If such a sequence element were located at a somewhat variable distance from the IES termini, it is unlikely that our analysis of non-random base composition would have detected it. In any event, the results presented provide a good indication that the IES ends represent at least part of the *cis*-acting sequences that are recognized by the cellular excision machinery. As such, the terminal IES sequences provide a logical starting point for attempting to identify the components of the excision machinery that directly interact with the IESs.

ACKNOWLEDGEMENTS

The authors wish to thank Drs L. Amar, J. Forney, E. Meyer, J. Preer and H. Schmidt for kindly providing sequence data prior to publication and/or for insightful discussions. This work was supported by National Science Foundation grant MCB-9414416 to L.A.K. and National Institutes of Health grant GM25203 to G.H.

REFERENCES

1 Klobutcher,L.A. and Prescott,D.M. (1986) In Gall, J. (ed.), *The Molecular Biology of Ciliated Protozoa*. Academic Press, NY, pp. 111–154.
2 Yao,M.-C. (1989) in Berg,D.E. and Howe,M.M. (eds), *Mobile DNA*. American Society for Microbiology, Washington, D.C., pp. 715–734.
3 Klobutcher,L.A. and Jahn,C.L. (1991) *Curr. Opin. Genet. Dev.*, **1**, 397–403.
4 Herrick,G. (1994) *Semin. Dev. Biol.*, **5**, 3–12.
5 Prescott,D.M. (1994) *Microbiol. Rev.*, **58**, 233–267.
6 Klobutcher,L.A., Jahn,C.L. and Prescott,D.M. (1984) *Cell*, **36**, 1045–1055.
7 Ribas-Aparicio,R.M., Sparkowski,J.J., Proulx,A.E., Mitchell,J.D. and Klobutcher,L.A. (1987) *Genes Dev.*, **1**, 323–336.
8 Herrick,G., Cartinhour,S.W., Williams,K.R. and Kotter,K.P. (1987) *J. Protozool.*, **34**, 429–434.
9 Eder,C., Maerker,C., Meyer,J. and Lipps, H.J. (1993) *Int. J. Dev. Biol.*, **37**, 473–477.
10 Bierbaum,P., Donhoff,P.B. and Klein,A. (1991) *Mol. Microbiol.*, **5**, 1567–1575.
11 Austerberry,C.F. and Yao,M.-C. (1988) *Mol. Cell Biol.*, **8**, 3947–3950.
12 Katoh,M., Hirono,M., Takemasa,T., Kimura,M. and Watanabe,T. (1993) *Nucleic Acids Res.*, **21**, 2409–2414.
13 Heinonen,T.Y.K. and Pearlman, R.E. (1994) *J. Biol. Chem.*, **269**, 17428–17433.
14 Wells,J.M., Ellingson,J.L.E., Catt,D.M., Berger,P.J. and Karrer,K.M. (1994) *Mol. Cell Biol.*, **14**, 5939–5949.
15 Godiska,R. and Yao,M.-C. (1990) *Cell*, **61**, 1237–1246.
16 Godiska,R., James,C. and Yao,M.-C. (1993) *Genes Dev.*, **7**, 2357–2365.
17 Baird,S.E., Fino,G.M., Tausta,S.L. and Klobutcher,L.A. (1989) *Mol. Cell. Biol.*, **9**, 3793–3807.
18 Jahn,C.L., Krikau,M.F. and Shyman,S. (1989) *Cell*, **59**, 1009–1018.
19 Krikau,M.F. and Jahn,C.L. (1991) *Mol. Cell. Biol.*, **11**, 4751–4759.
20 Herrick, G., Cartinhour,S., Dawson,D., Ang,D., Sheets,R., Lee,A. and Williams,K. (1985) *Cell*, **43**, 759–768.
21 Williams,K., Doak,T. G. and Herrick, G. (1993) *EMBO J.*, **12**, 4593–4601.
22 Mitchum,J.L., Lynn,A.J. and Prescott,D.M. (1992) *Genes Dev.*, **6**, 788–800.
23 Klobutcher,L.A. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 1979–1983.

24 Tausta,S.L., Turner,L.R., Buckley,L.K. and Klobutcher,L.A. (1991) *Nucleic Acids Res.*, **19**, 3229–3236.
25 Klobutcher,L.A., Turner,L.R. and LaPlante,J. (1993) *Genes Dev.*, **7**, 84–94.
26 Jaraczewski,J.W. and Jahn,C.L. (1993) *Genes Dev.*, **7**, 95–105.
27 Steele,C.J., Barkocy-Gallagher,G.A., Preer,L.B. and Preer,J.R. Jr (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 2255–2259.
28 Amar,L. (1994) *J. Mol. Biol.*, **236**, 421–426.
29 Scott, J., Leeck, C., and Forney, J. (1994) *Nucleic Acids Res.*, **22**, 5079–5084.
30 Ghosh,S., Jaraczewski,J.W., Klobutcher,L.A. and Jahn,C.L. (1994) *Nucleic Acids Res.*, **22**, 214–221.
31 Goodrich,J.A., Schwartz,M.L. and McClure,W.R. (1990) *Nucleic Acids Res.*, **18**, 4993–5000.
32 Gribskov,M., Luthy,R. and Eisenberg, D. (1990) *Methods Enzymol.*, **183**, 146–159.
33 Preer,J.R., Jr. (1986) In Gall,J. (ed.), *The Molecular Biology of Ciliated Protozoa*. Academic Press, NY, pp. 301–339.
34 Collins,J., Forbes,E. and Anderson,P. (1989) *Genetics*, **121**, 47–55.
35 Henikoff,S. (1992) *New Biol.*, **4**, 382–388.
36 Dreyfus,D.H. (1992) *Mol. Immunol.*, **29**, 807–810.
37 Radice,A.D., Bugaj,B., Fitch,D.H.A. and Emmons,S.W. (1994) *Mol. Gen. Genet.*, **244**, 606–612.
38 Henikoff,S. (1991) *New Biol.*, **3**, 1148–1154.
39 Doak,T.G., Doerder,F.P., Jahn,C.L. and Herrick,G. (1993) *Proc. Natl. Acad. Sci. USA*, **91**, 942–946
40 Jahn,C.L., Doktor,S.Z., Frels,J.S., Jaraczewski,J.W. and Krikau,M.F. (1993) *Gene*, **133**, 71–78.
41 Stoltzfus,A., Spencer,D.F., Zuker,M., Logsdonm,J.M., Jr and Doolittle,W.F. (1994) *Science*, **265**, 202–207.
42 Van Luenen,H.G.A.M., Colloms,S.D. and Plasterk,R.H.A. (1994) *Cell*, **79**, 293–301.
43 Knecht,K. and Klobutcher,L.A. (1995) *Eur. J. Protist.*, **31**, in press.
44 Robertson,H.M. (1993) *Nature*, **362**, 241–245.
45 Berg,D.E. and Howe,M.M. (1989) *Mobile DNA*. American Society for Microbiology, Washington, D.C.
46 Johnson,R. and Simon,M. (1985) *Cell*, **41**, 781–791.