Volume 13 Number 20 1985 Nucleic Acids Research

Sequence comparison of the rDNA introns from six different species of *Tetrahymena*

Henrik Nielsen and Jan Engberg

Department of Biochemistry B, Panum Institute, University of Copenhagen, Blegdamsvej 3C,
DK-2200 Copenhagen N, Denmark

Received 10 September 1985; Accepted 27 September 1985

ABSTRACT

We have studied the sequence variation of the rDNA intron among six species of Tetrahymena. From these data, the intron appears to be relatively well conserved in evolution. We have evaluated the sequence variations among the most distant of these species in relation to the secondary structure model for the intron RNA of Cech et al. (Proc. Natl. Acad. Sci. U.S.A. 80, 3903 (83)). Most of the sequence variation in the four new sequences reported here is found in single stranded loops in the model. However, in four cases we found nucleotide substitutions in duplex stem regions, two of them involving compensating base pair changes. Interestingly, one of these is found in a region that is known to be dispensable in the in vitro splicing reaction suggesting differences between the in vivo and in vitro reactions. One of the single nucleotide deletions is found in the socalled "internal guide sequence" which has been implicated in the alignment process during splicing. In conclusion, none of the observed natural sequence variations are in disfavor of the proposed secondary structure model.

INTRODUCTION

The discovery of the ability of Tetrahymena pre-rRNA to undergo selfsplicing in vitro (1) has led to considerable interest in the sequence and secondary structure of type I introns (2) and their flanking regions. Up till now, the sequence of the rDNA intron has been reported for two species of Tetrahymena, namely T. pigmentosa (3) and T. thermophila (4). A secondary structure model for the RNA has been proposed based on phylogenetic comparisons, base pairing rules (5,6,7) and RNase digestion studies of the T. thermophila intron (5). Subsequently, the effect of sequence modifications on the splicing reaction has been studied in vitro (8) as well as in vivo in E. coli cells transformed with plasmids carrying modified intron containing rDNA fragments (9,10). In the absence of a transformation system of Tetrahymena, the sequence modification approach is limited to the study of the splicing reaction in vitro or in a heterologous environment. As an alternative to this, we have studied the natural sequence variation among intron-containing Tetrahymena species. In a recent survey of the T. pyriformis complex based on restriction enzyme mapping of rDNA, we have found introns in four of the newly described species (11). We here report the DNA sequence of two of these, T. malaccensis and T. sonneborni as well as that of two previously described species, T. hyperangularis and T. cosmopolitanis. In the above mentioned study, we proposed that T. pigmentosa, T. sonneborni, T. hyperangularis and T. cosmopolitanis belong to one species cluster to which T. thermophila is distantly and T. malaccensis intermediately related. This view is supported by other studies on e.g. cytoskeletal proteins (12) and has recently been substantiated by sequence comparisons of the 17S rRNA genes (M. Sogin, pers. comm.). Thus, the present study includes closely as well as distantly related species within the T. pyriformis complex.

MATERIALS AND METHODS

Cell culture and DNA extraction. T. cosmopolitanis UM913, T. hyperangularis 10 IEN, T. malaccensis MP 75 and T. sonneborni XI6 were kindly provided by Dr. E. Simon, Univ. of Illinois at Urbana-Champaign. The strains were cultivated in a complex proteose peptone medium at 28°C. Different strategies were used for the isolation of rDNA for cloning. In the case of T. cosmopolitanis and T. hyperangularis, the rDNA was extracted by the hot phenol method (13) and purified on a CsCl density gradient. In the case of T. sonneborni, the CsCl gradient was omitted and finally, in the case of T. malaccensis, total DNA was extracted from a 1 ml culture by a simple miniprep procedure (14) and an intron-containing rDNA fragment partially purified by BamHl digestion and preparative agarose gel electrophoresis.

Cloning and DNA sequencing. In all cases, a 1.6 kb HindIII fragment containing the entire intron was isolated from the rDNA material by preparative electrophoresis on a low temperature gelling 1% agarose gel (Litex, Denmark) and ligated to HindIII digested pACYC 184 treated with alkaline phosphatase. The recombinant DNA was used to transform E. coli MC 1000 or R80 by the $CaCl_2$ transformation procedure (15). Large scale plasmid preparations were performed according to the alkaline lysis method (16). DNA sequencing was done by the Maxam and Gilbert method (17). Overlapping fragments on both strands were sequenced except at a few non-critical positions. All enzymes were purchased from Boehringer Mannheim.

RESULTS AND DISCUSSION

The sequences of the four introns sequenced in this study and those previously published are compiled in Figure 1. The sequences of T. hyperangularis, T. cosmopolitanis and T. sonneborni are identical and differ by only one nucleotide substitution to the sequence of T. pigmentosa. In comparisons between the more distant species, the sequences differ by both substitutions and deletions, the number of substitutions being approximately twice the number of deletions in all comparisons. It is difficult to evaluate the extent of the overall sequence divergence because comparable data from other parts of the rDNA molecule are scarce. The immediately flanking regions of the 26S rRNA gene are known to be extremely well conserved in evolution (18) and thus unsuited for comparison in this case. However, the sequence of the entire 17S rRNA gene has recently become available from several species (M. Sogin, pers. comm.) and the sequence of a segment of the central rDNA spacer has previously been reported by one of us (19). The difference of one in 407 nucleotides between T. pigmentosa and the three identical species represented by T. hyperangularis is very similar to the sequence divergence between the 17S rRNA genes of the same species. In contrast, the sequence divergence is approximaely ten times (3/141) higher in the central rDNA spacer. When T. pigmentosa and T. thermophila are compared, the sequence divergence in the intron is approximately four times higher than in the 17S rRNA gene, but much lower than in the central rDNA spacer in which the sequenced segment was found to be composed of a 38 bp block of complete homology and 100 bp of complete divergence. No comparable data is presently available from T. malaccensis. From the present data we suggest that the ribosomal RNA intron in Tetrahymena is evolving somewhat

```
tgacgcaattcaaccaagcgcgggtaaacggcgggagtaactatgactctctAAATAGCA   8  T. thermophila
                                                        T         T. pigmentosa
                                                        T         T. hyperangularis
                                                        T T       T. malaccensis

ATATTTACCTTTGGAGGGAAAAGTTATCAGGCATGCACCTGGTAGCTAGTCTTTAAACCA  68  T. thermophila
 A                              C        A                        T. pigmentosa
 A                              C        A                        T. hyperangularis
 A    -                         A        A                        T. malaccensis

ATAGATTGCATCGGTTTAAAAGGCAAGACCGTCAAATTGCGGGAAAGGGGTCAACAGCCG 128  T. thermophila
                   T                          A                   T. pigmentosa
                   T                          A                   T. hyperangularis
    C               T   G                     A                   T. malaccensis

TTCAGTACCAAGTCTCAGGGGAAACTTTGAGATGGCCTTGCAAAGGGTATGGTAATAAGC 188  T. thermophila
                                              A                   T. pigmentosa
                                              A                   T. hyperangularis
                GA                            A                   T. malaccensis

TGACGGACATGGTCCTAACCACGCAGCCAAGTCCTAAGTCAACAGATCTTC--TGTTGAT 248  T. thermophila
         G                                  TT  -GG--             T. pigmentosa
         G                                  TT  -GG--             T. hyperangularis
         G                               G  CTCT G AG  C          T. malaccensis

ATGGATGCAGTTCACAGACTAAATGTCGGTCGGGGAAG-AT--GTATTCTTCTCATAAGA 308  T. thermophila
                                      A  AG                       T. pigmentosa
                                      A  AG                       T. hyperangularis
     C                              G - A-- -                     T. malaccensis

TATAGTCGGACCTCTCCTTAATGGGAGCTAGCGGATGAAGTGATGCAACACTGGAGCCGC 368  T. thermophila
                             G            G        CA             T. pigmentosa
                             G            G         A             T. hyperangularis
                             G T   A      G         A     A       T. malaccensis

TGGGAACTAATTTGTATGCGAAAGTATATTGATTAGTTTTGGAGTACTCGtaaggtagcc 418  T. thermophila
C         -- -A  ---   -C   A -            C  A                   T. pigmentosa
C         -- -A  ---   -C   A -            C  A                   T. hyperangularis
C         -- A -   --  -- - - -            C  A                   T. malaccensis


aaatgcctcgtcatctaattagtgacgcgcatgaatggatta   - intron: 413 bp -   T. thermophila
                                             - intron: 407 bp -   T. pigmentosa
                                             - intron: 407 bp -   T. hyperangularis
                                             - intron: 403 bp -   T. malaccensis
```

Fig. 1. Compilation of the DNA sequence of the ribosomal RNA intron from various species of Tetrahymena. The flanking regions of 26S rRNA gene sequence is shown in lower case letters. The numbers refer to a hypothetical intron (upper case letters) of 418 bp in order to accommodate the length variation among the introns from the different species. The underlined sequences are those considered in Figs. 2-5. The sequences of T. cosmopolitanis and T. sonneborni are identical to T. hyperangularis. T. thermophila was from (4) and T. pigmentosa from (3).

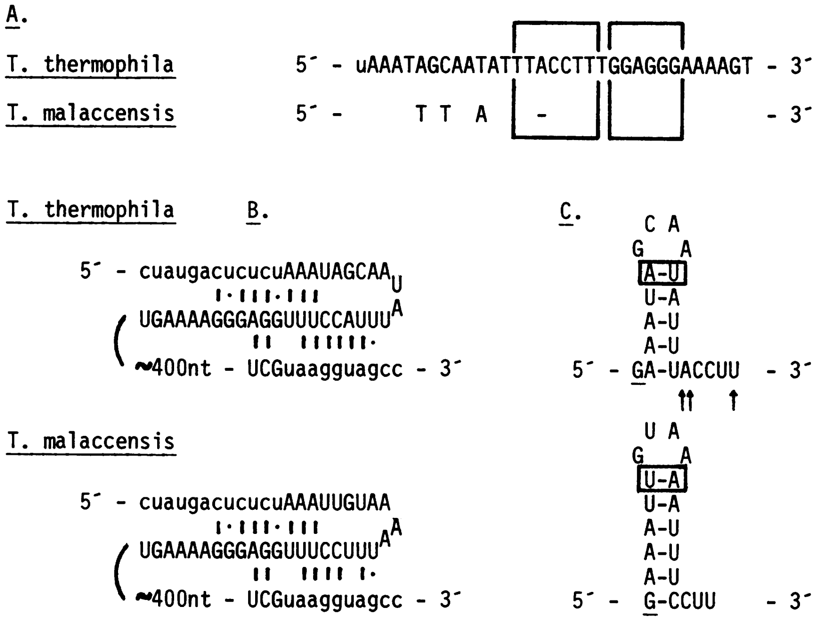

Fig. 2. A. Sequence comparison of the 5'-end of the introns of T. thermophila and T. malaccensis with emphasis on the deletion of A_{15} in T. malaccensis. The boxed regions are capable of basepairing with the flanking exons (lower case letters) in the pre-rRNA as shown in B. In C. is shown a potential secondary structure of the 5'-end of the excised intron. The boxed basepair is in this model a compensating pair of transversions between T. thermophila and the other species. The underlined G is added during excision of the intron. Following excision, the intron can undergo autocyclization at the major (double arrow) or minor (single arrow) cyclization sites.

faster than the rRNA coding regions but considerably slower than the central rDNA spacer.

The individual sequence differences between the introns can be evaluated not only in an evolutionary context but also in a functional context. In the following, we try to relate the sequence variation in four distinct regions to the secondary structure model of Cech et al. (5).

Position 15 (Fig. 2). Compared to the other species, the A at position 15 is deleted in T. malaccensis. This deletion is apparently not critical for splicing since the strain in which the intron containing allele was found is homogenously intron^{+} and therefore have to splice its pre-rRNA in order to produce

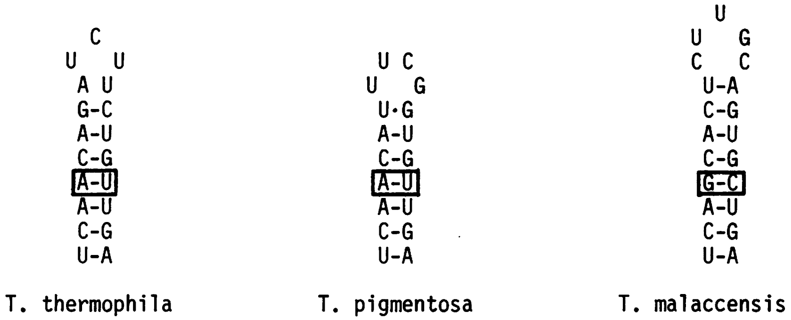

Fig. 3. Potential secondary structure according to (5) of segment 227-247. The boxed nucleotides are those involved in a compensating pair of transitions between T. malaccensis and the other species.

mature 26S rRNA. Nevertheless, A_{15} is believed to be involved in two presumably important aspects of RNA splicing: the formation of the precise alignment structure (20) and the autocyclization of the excised linear intron (21). Davies et al. (20) originally found, that a sequence element in type I introns has the ability to form basepairs with the immediately flanking parts of the two exons thereby aligning the exons properly for ligation after intron excision (Figure 2). Pleij et al. (22) has further extended the implication of this structure in splicing by suggesting that it, together with an additional stem located further downstream, is capable of transition between two states by the formation of a coaxial, quasi-continous stem. The deletion in T. malaccensis reduces the basepairing between the intron and the 3'-exon with one basepair. The significance of this in relation to the above mentioned model is unknown. It should be emphasized, that the existence of the alignment structure has not been directly demonstrated. In fact, it has recently been shown, that an RNA splicing process can occur in vitro in a RNA molecule where the intron-exon junctions have been removed (J. Price et al., pers. comm.). Thus, the alignment structure does not seem to play a role in the autocatalytic process itself but rather in the selection of splice sites.

Following excision, the linear intron is capable of autocyclization at positions A_{15} (major) or C_{19} (minor) with the

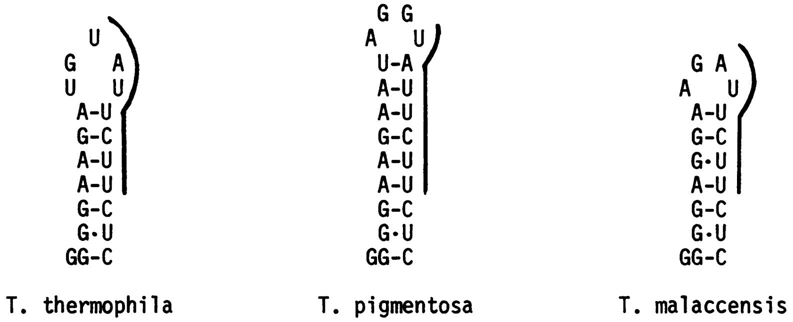

Fig. 4. Potential secondary structure according to (5) of segment 279-302. The sequence homologous with the yeast mitochondrial 3'-end processig signal is underlined. The stem containing this sequence is located very close to the central core of the intron RNA. At the 5'-end it is flanked by the socalled box 9R element (2) and likewise it is immediately followed at the 3'-end by the conserved box 2 sequence element.

subsequent release of 15 or 19 nucleotides (21) (including the externally added G). The deletion of A_{15} in T. malaccensis allows for a basepairing between the externally added G and the C in position 15 in T. malaccensis in the potential 5'-stem/loop structure (Figure 2C). It would be interesting to test if this change directs autocyclization to the minor cyclization site in T. malaccensis.

Position 227-247 (Fig. 3). This segment corresponds in the secondary structure model to the "f" stem and loop (5). This structure has been shown to be dispensable for in vitro splicing (8) and splicing in E. coli (9). In both types of experiments, a series of Bal 31 deletions starting at the Bgl II site was tested and it was concluded that deletion of the entire stem was without effect on splicing. It was therefore surprising to us to find that the sequence variation between T. malaccensis and the other species involved a compensating pair of transitions in the "f" stem. This type of double mutations is usually taken as an indication of the existence of the structure at the secondary structure level. If the "f" stem is existing and conserved, it is reasonable to assume that it plays a role in splicing which would suggest a difference between the in vitro and in vivo splicing reactions. An involvement of protein binding has been speculated

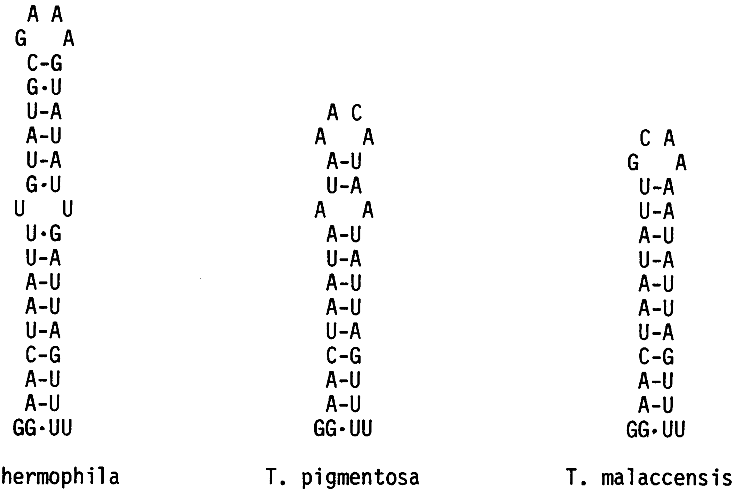

Fig. 5. Potential secondary structure according to (5) of segment 371-408.

on earlier (23) and could account for part of the observed differences in the splicing rate between in vivo and in vitro splicing.

Position 279-302 (Fig. 4). The loop structure of this segment is highly variable whereas the stem (the "h" stem in (5)) is well conserved perhaps because of its proximity to the central core of the overall intron structure. The central core is formed mainly by the sequence elements characteristically conserved among type I introns. In most of the cases where a type I intron contains a region with an open reading frame, this region is found at a position equivalent to the position of the "h" stem in the intron in Tetrahymena. We have previously speculated that the variability in the loop at this characteristic position was reminiscent of a deletion event perhaps leading to the loss of a coding region in Tetrahymena during the course of evolution (14). Now we find supporting evidence for this hypothesis in the fact that the sequence 5' - UAUUCUU - 3' found in the stem at position 293-299 is exactly the last seven nucleotides of the dodecamer sequence 5' - AAUAAUAUUCUU - 3' found at the 3' - end of yeast mitochondrial reading frames and believed to be a processing signal (24). We suggest that the heptanucleotide found in

Tetrahymena is a truncated version of this signal remaining after the deletion of the open reading frame region itself. The homology can be extended when comparison is made with one particular yeast open reading frame, namely the unassigned reading frame URF2 (26). In this case, the homology starting at the processing signal and extending downstream involves 20 matches and 3 deletions on the URF2 side. The homoloy extends over the conserved sequence element box 2 (2) with one of the yeast deletions in the middle of this sequence. Admittedly, the segments are very AT-rich (16/23 and 16/20) and it is possible that this extended homology is purely coincidental.

The sequence of T. malaccensis contains a "new" HinfI site at a position corresponding to the end of the stem structure. This provides a convenient possibility for perturbation of the intron structure close to the central core by insertion and deletion studies as in (8) in order to obtain direct evidence for the involvement of this part of the molecule in splicing.

Position 371-408 (Fig. 5). The sequence in this region is the most variable in the intron. In spite of this, it is possible in all species to form the "k" stem (5) at the secondary structure level. At present, no specific feature has been ascribed to this structure.

CONCLUSION

In conclusion, the sequence data contributed by this study is in agreement with the secondary structure model of Cech et al. (5). The favorable combination of a considerable amount of structural data and a detailed knowledge of the splicing process (26) should stimulate further work on several questions for example on the origin of the intron by taking advantage of the fact, that intron$^+$ and intron$^-$ rDNA is found within the same genus and even within the same species (the highly polymorphic T. pigmentosa and T. malaccensis (11,27,28)). If indeed, the splicing process in vivo, as is the case in vitro, occurs independent of other informational structure, it should be possible to deduce its evolutionary history by comparison of variation in sequence among different species in the intron and e.g. in the 17S rRNA gene. More specifically, it should be possible to tell if the observed length differences among the introns are due to deletions or insertions. One other question that needs further consideration is the observation of sequence elements in Tetrahymena homologous with, and functionally related to yeast mitochondrial sequences. Such a homology was observed long ago (2,20) for the sequence elements involved in selfsplicing of type I introns and is here proposed for certain cases of 3'-end processing of mRNA's. Further work involving comparative sequence data will be necessary in elucidating whether this relationship between the rDNA intron in Tetrahymena and certain mitochondrial introns is due to a common origin of the genomes or secondarily aquired e.g. by transposable elements.

ACKNOWLEDGEMENTS

We are grateful to Prof. O. Nanney and Dr. E. Simon for supplying the different Tetrahymena strain, to Mr. F. Frenzel and Ms. A. Iversen for technical assistance and to Mrs. E. Høyer for typing the manuscript. This work was supported in part by the Danish Natural Science Research Council.

REFERENCES

1. Cech, T.R., Zaug, A.J. and Grabowski, P.J. (1981) Cell 27, 487-496.
2. Michel, F., Jacquier, A. and Dujon, B. (1982) Biochimie 64, 867-881.
3. Wild, M.A. and Sommer, R. (1980) Nature 283, 693-694.
4. Kan, N.C. and Gall, J.G. (1982) Nucl. Acids Res. 10, 2809-2822.
5. Cech, T.R., Tanner, N.K., Tinoco, I., Weir, B.R., Zuker, M. and Perlman, P.S. (1983) Proc. Natl. Acad. Sci. 80, 3903-3907.
6. Michel, F. and Dujon, B. (1983) EMBO J. 2, 33-38.
7. Waring, R.B., Scazzocchio, C., Brown, T.A. and Davies, R.W. (1983) J. Mol. Biol. 167, 595-605.
8. Price, J.V., Kieft, G.L., Kent, J.R., Sievers, E.L. and Cech, T.R. (1985) Nucl. Acids Res. 13, 1871-1889.
9. Price, J.V. and Cech, T.R. (1985) Science 228, 719-722.
10. Waring, R.B., Ray, J.A., Edwards, S.W., Scazzocchio, C. and Davies, R.W. (1985) Cell 40, 371-380.
11. Nielsen, H., Simon, E.M. and Engberg, J. (1985) J. Protozoology 32, 480-485.
12. Williams, N.E., Buhse, H.E. and Smith, M.G. (1984) J. Protozoology 31, 313-321.
13. Engberg, J., Din, N., Eckert, W.A., Kaffenberger, W. and Pearlman, R.E. (1979) J. Mol. Biol. 142, 289-313.
14. Nielsen, H. and Engberg, J. (1985) Biochim. Biophys. Acta

825, 30-38.
15. Mandel, M. and Higa, A. (1970) J. Mol. Biol. 53, 154-
16. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning. A Laboratory Handbook. Cold Spring Harbor Laboratory.
17. Maxam, A.M. and Gilbert, W. (1980) Meth. Enzymol. 65, 499-560.
18. Gerbi, S.A., Gourse, R.L. and Clark, C.G. (1982) Cell Nucl. 10, 351-386.
19. Engberg, J. (1983) Nucl. Acids Res. 11, 4939-4945.
20. Davies, R.W., Waring, R.B., Ray, J.A., Brown, T.A. and Scazzocchio, C. (1982) Nature 300, 719-724.
21. Grabowski, P.J., Zaug, A.J. and Cech, T.R. (1981) Cell 23, 467-476.
22. Pleij, C.W.A., Rietveld, K. and Bosch, L. (1985) Nucl. Acids Res. 13, 1717-1731.
23. Cech, T.R. (1985) Int. Rev. Cytol. 93, 3-22.
24. Osinga, K.A., De Vries, E., Van der Horst, G. and Tabak, H.F. (1984) EMBO J. 3, 829-834.
25. Macino, G. and Tzagoloff, A. (1980) Cell 20, 507-517.
26. Cech, T.R. (1985) in Gall, J.G. (ed.) The Molecular Biology of Ciliated Protozoa. Academic press, New York (in press).
27. Din, N. and Engberg, J. (1979) J. Mol. Biol. 134, 555-574.
28. Wild, M.A. and Gall, J.G. (1979) Cell 16, 565-573.