The intervening sequence of the ribosomal RNA gene is highly conserved between two *Tetrahymena* species

Nancy C.Kan* and Joseph G.Gall[†]

*Laboratory of Molecular Oncology, National Cancer Institute, Bethesda, MD 20205, and [†]Department of Biology, Yale University, New Haven, CT 06511, USA

Received 3 March 1982; Accepted 12 April 1982

ABSTRACT

The entire intervening sequence of *Tetrahymena thermophila* ribosomal DNA has been determined. It is 413 nucleotides long and has the same splice junctions as those in *T. pigmentosa*. There is 93% homology between the intervening sequences in the two species, and 100% homology between their adjacent 26S RNA coding regions.

INTRODUCTION

An intervening sequence (or intron) in the macronuclear ribosomal RNA gene of *Tetrahymena pigmentosa* was first observed by Wild and Gall (1). It is located near the 3'-end of the 26S coding region of the extrachromosomal rDNA, and its entire sequence of 407 nucleotides has been determined (2). Since the intervening sequence is only present in certain strains of *T. pigmentosa*, Wild and Sommer (2) were able to identify the splice junctions by comparing the nucleotide sequences containing and lacking the intervening sequence.

Most studies on the transcription and processing of *Tetrahymena* ribosomal genes have been done in the species *T. thermophila*, whose extrachromosomal rDNA contains an intervening sequence similar in size and location to that from *pigmentosa* (3,4). Hybridization experiments indicate that the intervening sequence is transcribed as part of a large stable RNA (3, 4), and its excision from the precursor is the first step during ribosomal RNA processing (5). Investigations of RNA synthesis in isolated nuclei reveal that the excised intervening sequence exists as an intact linear molecule (6). The same *in vitro* system also catalyzes the addition of a guanosine residue to the 5'-end of the intervening sequence during excision, and the intervening sequence is subsequently converted to a covalently closed circle (7, 8). These observations suggest a specialized and significant function of the intervening sequence in its own excision.

On the other hand, analysis of rDNA structure indicates that some species and strains of Tetrahymena contain no intervening sequence in their extrachromosomal rDNA (9, 1). Since the extrachromosomal rDNA molecules are homogeneous in each clonal population of a species, the presence or absence of the intervening sequence does not appear to affect ribosomal RNA transcription. As part of studies on the processing of rRNA molecules in Tetrahymena, we report in the present work the primary structure of the intervening sequence in T. thermophila, and the determination of its boundaries with the coding region. This also allows us to compare the two intervening sequences in T. thermophila and T. pigmentosa.

MATERIALS AND METHODS

Preparations of plasmid DNA and extrachromosomal rDNA

The restriction enzyme Hind III cleaves T. thermophila extrachromosomal rDNA into seven fragments. The 1.6 kb fragment containing the intervening sequence was inserted into the bacterial plasmid pBR313 by Ed Stephenson. This recombinant plasmid is designated pTt116. Restriction enzyme mapping indicates that pTt116 contains 2 copies of the 1.6 kb fragment arranged in a head-to-tail fashion. The 3'-end of the 26S coding region is located near the Eco RI site of pBR313, while the 5'-end is located near the Bam HI site of pBR313 (see Figure 3A).

The pTt116 plasmid DNA was prepared essentially by the method of Guerry *et al.* (10). T. thermophila cells were grown and its extrachromosomal rDNA was extracted as described by Wild and Gall (1). Restriction fragments were purified from low-gelling-temperature agarose gels (Marine Colloids, Inc.) by melting the gel pieces at 65°C, 15 min. in the presence of an equal volume of phenol. The DNA solution was extracted with phenol and precipitated with ethanol. DNA fragments from polyacrylamide gels were eluted by crushing the gel pieces and soaking them in 10mM Tris-HCl, pH 8, 1mM EDTA, and 100mM NaCl at 42°C overnight. The mixture was centrifuged through a fiber glass filter purchased from Isolab, Inc. The DNA solution was then extracted with phenol and precipitated with ethanol.

End-labeling and sequencing of DNA

Restriction fragments were labeled either at their 3'-ends using E. coli DNA polymerase I large fragment and [α -³²P]dNTP (11), or at their 5'-ends using T4 polynucleotide kinase and [γ -³²P]ATP (12). [α -³²P]dNTP (400 Ci/mmol or 2,000 Ci/mmol) and [γ -³²P]ATP (about

2,000 Ci/mmol) were purchased from Amersham or New England Nuclear. *E. coli* DNA polymerase I large fragment was from Boehringer Mannheim Corporation, and T4 polynucleotide kinase (about 20 U/ μ l) from P-L Biochemicals, Inc.

Mapping of restriction sites on pTt116 was carried out by digesting either the plasmid DNA or end-labeled restriction fragments with different restriction enzymes. Since the nucleotide sequence of the entire 1.6 kb insert was determined (N. Kan and J. Gall, manuscript in preparation), all the restriction sites determined by digesting DNA were confirmed by the sequence data. Sequencing of DNA was carried out as described by Maxam and Gilbert (12, 13). A restriction enzyme map and sequencing strategy of the rDNA intron are shown in Figure 1.

S1 nuclease mapping of intron-exon junctions

The procedure is essentially as described by Din *et al* (3) with the following modifications. Hybridization with either 26S ribosomal RNA or total RNA of *T. thermophila* kindly provided by Jan Engberg was carried out at 42°C for at least 5 hours. S1 nuclease digestion was done at 42°C for 30 min. S1 nuclease was prepared by Clifford Yen. One unit is defined as the amount of S1 nuclease that digests 10 μ g of DNA at 37°C in 1 hour.

Computer analysis of DNA sequences

DNA sequences were searched for restriction enzyme sites, amino acid

Figure 1 Comparison of the restriction maps of *thermophila* and *pigmentosa* rDNA introns. The intron is represented by a solid line between two crosshatched bars which represent the 26S coding region. Transcription is from right to left. The symbol $\frown$ indicates the heterogeneous region in the cloned intron. The restriction enzymes shown in this figure are A = Alu I, B = Bgl II, F = Hinf I, H = Hha I, P = Hpa II, and Z = Hae III. Arrows above the restriction maps show the sequencing strategy used to determine the nucleotide sequence of *thermophila* rDNA intron. Small vertical lines indicate the restriction sites which were labeled for sequencing. Dotted lines represent the sequences which were not determined by gels.

sequence and secondary structure using the program written by Queen and Korn (13). The dot-plot comparison was kindly provided by Clyde Hutchison using a program written for microcomputers (C. Hutchison, R. Padgett, S. Weaver, C. Voliva, C. Jahn, and M. Edgell, manuscript in preparation; 14).

RESULTS

Nucleotide sequence of the 26S rRNA gene intron of *T. thermophila*

The 26S rRNA gene intron of *T. thermophila* has been localized within the 1.6 kb-Hind III fragment of the extrachromosomal rDNA (3,4,9). A detailed transcription map indicates that the 400 bp intron resides between the two Hha I sites (5, Figure 1). The entire 1.6 kb fragment purified from plasmid pTt116 has been sequenced (N. Kan and J. Gall, manuscript in preparation). Only the intron and its adjacent coding region are presented in this report (see Figures 1 and 5).

When the Bgl II-Hha I fragment from positions 232 to -32 (Figure 5) was labeled at the Bgl II site and then sequenced, the band pattern was unambiguous from positions 232 (the Bgl II end) to 174. Starting from position 173, it appeared that two bands could be assigned to one position (Figure 2). More careful analysis by gel electrophoresis revealed that this fragment contained two fragments very close in size. To investigate whether one of the fragments was a contaminant from other regions of rDNA or the bacterial vector pBR313, both fragments were isolated and sequenced individually. The result indicates that except for the region between positions 163 and 173 (Figure 5), the rest of the sequences are identical. One fragment is 7 nucleotides longer than the other one, and the sequences within the heterogeneous region are:

162 -GTCGCTTGTAGTAACGAATG- 174

162 -GGC CTTGCA AAGG- 174

Space is inserted into the shorter sequence to maximize homology and minimize gaps. The Hae III site at position 162 (underlined nucleotides) of the shorter fragment is not present in the longer fragment due to the sequence heterogeneity.

Two questions were then asked: (a) Since pTt116 contains two copies of the 1.6 kb fragment, did the heterogeneity occur in adjacent copies on the same plasmid, or were there two plasmids, each containing two 1.6 kb fragments of identical sequence? More importantly, (b) was the rDNA in the organism itself heterogeneous? If not, which sequence represented the intron in the rDNA?

Figure 2 The heterogeneous region in pTt116 is revealed by a sequencing gel. The 1.6 kb-Hind III fragment of *thermophila* extrachromosomal rDNA was purified from pTt116 and cleaved with Bgl II. The small Hind III-Bgl II (see Figure 3A) containing the 5'-end of the intron was isolated and labelled at its 3'-ends. The labeled fragment was again cleaved with Hha I, and the Bgl II-Hha I fragment from positions 232 to -32 was sequenced from the Bgl II end. The sequencing gels reveal an unambiguous band pattern from the Bgl II end at position 232 to position 174. Only part of an 8% gel is shown here. At position 173, a T co-migrates with a G, and again at position 167, a T co-migrates with an A. Other sequencing experiments (see text) demonstrate that the heterogeneous region in pTt116 is from position 173 to position 163.

Figure 3A demonstrates the strategy for answering these questions. pTt116 plasmid DNA was cleaved with restriction enzymes Bam HI and Bgl II. Fragments I and II were isolated and labeled at both ends. In order to detect a size difference of 7 nucleotides, a third restriction enzyme (Hha I in Figure 3A) which cleaves near the Bgl II site was used to digest

Figure 3 *Thermophila* rDNA intron is homogeneous in its sequence. (A) Solid lines represent the 1.6 kb-Hind III fragment of *thermophila* extrachromosomal rDNA, while dotted lines represent bacterial vector pBR313. Solid bars represent the rDNA intron. Long arrows indicate transcription direction. The symbol ∇ marks the heterogeneous region found by DNA sequencing. Asterisks indicate labeled ends of DNA fragments. (B) 5% polyacrylamide gels of (a) fragment I cleaved with Hha I; (b) fragment II cleaved with Hha I; (c) the Bgl II-Hind III fragment (500 bp) of *thermophila* extrachromosomal DNA cleaved with Hha I; (d) fragment I cleaved with Hae III; (e) fragment II cleaved with Hae III.

fragments I and II. If the heterogeneity was located in the same plasmid, either I or II would be homogeneous but different from each other in 7 nucleotides (271 and 264 bp). If a second plasmid existed, both I and II would be heterogeneous and give an identical pattern of a doublet. Figure 3B, lanes a and b demonstrate that the recombinant plasmid pTt116 contains two 1.6 kb fragments of different lengths. To answer the second question, extrachromosomal rDNA extracted from one clonal population of thermophila was cleaved with Hind III. The 1.6 kb fragment was again cleaved with Bgl II and labeled only at the Bgl II site using dGTP. The small Bgl II-Hind III fragment containing the heterogeneous region was purified and cleaved with Hha I. Figure 3B, lane c shows that a homogeneous band 264 bp in length was produced. Furthermore, fragment I does not have a Hae III site in the heterogeneous region (lane d), while both fragment II (lane e) and rDNA (data not shown) are readily cleaved by Hae III within this region. Thus the extrachromosomal rDNA purified from one clonal population is homogeneous in the intron, and its nucleotide sequence is reflected in fragment II.

S1 nuclease mapping of the intron-exon junctions

Figure 4A demonstrates the S1 nuclease experiment that defines the 5'-end of the intron-exon junctions. The Dde I-Dde I restriction fragment containing the 5'-boundary region was isolated. It starts at position 142 (see Figure 5) and extends to another Dde I site 300 bp upstream. This fragment was labeled at its 3'-ends and after strand separation, hybridized to mature ribosomal RNA. Only the DNA complementary to rRNA was protected from S1 nuclease digestion. However, the intron located at the 5'-end of the DNA strand was not protected by mature rRNA and was thus removed by S1 nuclease. Digestion did not affect the labeled 3'-end of the DNA strand. In order to determine the nucleotides at the splice junction, the same DNA strand was also subjected to the chemical cleavage reactions of Maxam and Gilbert (12), and the products were electrophoresed in parallel with the rRNA-protected fragment on a sequencing gel. Since the S1-treated fragment runs one and a half nucleotides slower than the same fragment after the chemical cleavage reactions (16), the most prominent bands in both S1 lanes indicate that the A (T in the RNA-like strand of the DNA) at position 4 is the last nucleotide at the 5'-end of the protected DNA strand. This result does not agree with the experiment which determined the 5'-junction by sequencing the ends of the intron transcript (A. Zaugg and T. Cech, see accompanying paper). RNA sequencing demonstrated that the A at position 1

Figure 4 S1 nuclease mapping of intron-exon junctions. (A) The 309 bp Dde I-Dde I fragment containing the 5'-half of the intron (from positions 1 to 142) and its adjacent coding region (from positions -1 to -167) was used in this experiment. After 3'-end labeling and strand separation, the DNA complementary to rRNA was hybridized with either 26S rRNA (the left S1 lane) or total RNA (the right S1 lane). Two units of S1 nuclease were used in both reactions. Since the intron-exon junction is 167 bases away from the labeled end, the fragments produced from the chemical cleavage reactions are not very well resolved on the 8% sequencing gel shown here. However, the DNA sequence in this region has been confirmed by other experiments using different restriction fragments. (B) The Bgl II-Hha I fragment from positions 232 to 452 was kinased at its 5'-ends and strand separated. The DNA strand with its label at the Hha I site was used to hybridize with total RNA. Either 20 units (the leftmost S1 lane) or 40 units of S1 nuclease were used. The rRNA-protected fragments and DNA were sequencing reactions electrophoresed on a 8% sequencing gel.

(see Figure 5) was the first nucleotide of the intron, which is the same as the 5'-junction in *pigmentosa* rDNA. It is very likely that the low S1 concentration used in the mapping experiment was not sufficient to remove all the single-stranded portion of the rRNA-protected fragment.

A similar experiment defining the 3'-end of the intron-exon junctions is shown in Figure 4B. The Bgl II-Hha I fragment from positions 232 to 452

in thermophila as in pimentosa.

Comparison of the introns in T. thermophila and T. pimentosa

The nucleotide sequence of the intron and adjacent coding region of thermophila rDNA, as well as its comparison with that of pimentosa rDNA are presented in Figures 5 and 6. The intron of thermophila rDNA consists of 413 nucleotides, 6 nucleotides longer than that of pimentosa rDNA. The nucleotide sequences demonstrate 93% homology between the two introns, and

Figure 6 Sequence homology demonstrated by a computer program. The template sequence (horizontal axis) represents the pimentosa rDNA intron of 407 nucleotides, while the argument sequence (vertical axis) represents the thermophila rDNA intron of 413 nucleotides. A stretch of 5 nucleotides is compared each time, and when a perfect match is detected by the computer, a dot is plotted on the diagram. The computer printout was kindly provided by Clyde Hutchison using a program written for microcomputers (C. Hutchison, R. Padgett, S. Weaver, C. Voliva, C. Jahn and M. Edgell, manuscript in preparation; 15).

100% homology between their adjacent coding regions. Most changes in the intron are single-base substitutions, which include 8 transitions and 14 transversions. There are three longer non-homologous regions, all located in the 3'-half of the intron. Inverted repeats are observed surrounding these regions. As shown in Figure 6, the two introns are highly conserved, their 5'-ends having a higher degree of sequence homology than the 3'-ends. The intron-exon junctions are identical between the two species. The mutations closest to the splice points are single-base substitutions at position 5 (the 5'-end) and position 410 (the 3'-end). The longest open reading frame in the thermophila intron is from positions 65 to 265, and could possibly code for a protein 67 amino acids long.

DISCUSSION

Three methods have been used to determine the intron-exon junctions of Tetrahymena extrachromosomal rDNA. The splice point of pigmentosa rDNA was defined by comparing two strains containing (6UM) and lacking (8ALP) the intervening sequence (2). However, if there was a small intervening sequence in 8ALP rDNA that had gone undetected, such a comparison would misplace the splice point within a short distance. The S1 nuclease method developed by Berk and Sharp (17) was thus used to map the intron-exon junctions of thermophila rDNA. The shortcoming of this method is that sometimes a series of three to four bands each separated by one nucleotide is produced. Because of these partial digestions at the end of the RNA-protected fragment, the location of the exact junction is ambiguous. When the most prominent bands produced from the highest S1 concentrations were chosen to represent the intron-exon junctions, both methods independently determined the 3'-junctions in thermophila and pigmentosa rDNA at the same position. At the 5'-junctions, the S1 nuclease mapping placed the first nucleotide of the intervening sequence four bases away from that of pigmentosa. The third method, sequencing the ends of the excised intron transcript (A. Zaug and T. Cech, the accompanying paper), however clearly demonstrated that they had the same junction. Since it is necessary to use a large excess of S1 nuclease to remove all the single-stranded DNA, the reason that the S1 mapping disagrees with the RNA sequencing data is probably due to low S1 concentrations.

All the intervening sequences found in Tetrahymena so far are about 400 bp long and located near the 3'-end of the 26S rRNA coding region (1,3,4,9). Two of them (thermophila and pigmentosa) have been sequenced

and they are highly conserved (93% homology). Although it is difficult to estimate how closely related these two species are evolutionarily, they were placed in different species groups according to the restriction patterns of their extrachromosomal rDNA (9). This grouping was reinforced by sequence comparison in the central region of the extrachromosomal rDNA among thermophila, pigmentosa and hyperangularis (18). These observations have led us to conclude that the intervening sequence is evolutionarily conserved among Tetrahymena species.

Since the intervening sequence is transcribed as part of the precursor rRNA in Tetrahymena (3, 4), we have observed within the rDNA a higher level of sequence homology in the intervening sequence than in the central region (18), which does not encode ribosomal RNA (5). However, the fact that the intervening sequence is almost as conserved as the rRNA coding region or probably even more so (N. Kan and J. Gall, manuscript in preparation) is puzzling. In many eucaryotic genes transcribed by polymerase II, mutations occur more frequently in the intervening sequence than in the coding region, presumably because the intervening sequence is not represented in the final product, although its location and size within a gene are relatively stable among species (19). In the case of Drosophila ribosomal DNA, it is known that the rDNA repeats containing an intervening sequence are not transcribed (20). While the coding region of these repeats, as well as the location of the intervening sequence are highly conserved between D. melanogaster and D. virilis, the intervening sequence itself displays great sequence divergence (21, 22). Therefore, the sequence conservation in Tetrahymena may indicate a specialized and significant function of the intervening sequence, perhaps a function in its own excision. The lack of sequence homology at the splice junctions of the intervening sequences transcribed by polymerases I and II suggests that they are recognized by different components during excision. Furthermore, studies in isolated nuclei of Tetrahymena reveal a completely different splicing mechanism from that involved in the splicing of yeast tRNA precursors (8).

Although the intervening sequence is evolutionarily stable and probably functionally important, some strains and species of Tetrahymena have no intervening sequence in their extrachromosomal rDNA (1,5). Phenotypically, these organisms are as normal as the ones containing the intervening sequence. One possible explanation is that there are two completely different systems of rRNA maturation in Tetrahymena. One system

requires splicing to produce functional rRNA, while the other does not. Studies on the mechanism of rRNA maturation in species containing and lacking the intervening sequence could provide useful information on this question.

Finally, the two different 1.6 kb inserts of thermophila extrachromosomal rDNA found in pTt116 may represent real sequence heterogeneity in the micronuclear rDNA, and the longer insert is derived from a minor fraction of the extrachromosomal rDNA molecules in a clonal population. The other possibility is that the heterogeneous region in pTt116 is an example of sequence rearrangement in E. coli. If this is the case, we find it difficult to explain why only one copy of the inserts was rearranged while the other one maintains the same sequence as the thermophila rDNA intron.

ACKNOWLEDGEMENTS

We are indebted to Jan Engberg for his gift of Tetrahymena 26S and total ribosomal RNA, Ed Stephenson for the gift of recombinant plasmid pTt116, Clyde Hutchison for providing Figure 6, and Clifford Yen for his S1 nuclease. We also thank Tom Cech, Art Zaugg, Nanni Din, Martha Wild and Jules O'Rear for useful discussions. This work was supported by a National Institutes of Health postdoctoral fellowship GM 07653 to N.C.K., and an American Cancer Society research grant CD 37B to J.G.G.

REFERENCES

- 1 Wild, M.A. and Gall, J.G. (1979). *Cell* **16**, 565-573.
- 2 Wild, M.A. and Sommer, R. (1980). *Nature* **283**, 693-694.
- 3 Din, N., Engberg, J., Kaffenberger, W. and Eckert, W.A. (1979). *Cell* **18**, 525-532.
- 4 Cech, T.R. and Rio, D.C. (1979). *Proc. Natl. Acad. Sci. U.S.A.* **76**, 5051-5055.
- 5 Engberg, J., Din, N., Eckert, W.A., Kaffenberger, W. and Pearlman, R.E. *J. Mol. Biol.* (1980). **142**, 289-313.
- 6 Zaugg, A.J. and Cech, T.R. (1980). *Cell* **19**, 331-338.
- 7 Grabowski, P.J., Zaugg, A.J. and Cech, T.R. (1981). *Cell* **23**, 467-476.
- 8 Cech, T., Zang, A. and Grabowski, P. (1982). *Cell*, in press.
- 9 Din, N. and Engberg, J. (1979). *J. Mol. Biol.* **134**, 555-574.
- 10 Guerry, P., LeBlanc, D.J. and Falkow, S. (1973). *J. Bacteriol.* **116**, 1064-1066.
- 11 Sanger, F., Nicklen, S. and Coulson, A.R. (1977). *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463-5467.
- 12 Maxam, A.M. and Gilbert, W. (1977). *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560-564.
- 13 Maxam, A.M. and Gilbert, W. (1980). *Methods in Enzymology* **65**, 499-560.
- 14 Queen, C.L. and Korn, L.J. (1980). *Methods in Enzymology* **65**, 595-609.

- 15 Edgell, M.H., Weaver, S., Jahn, C.L., Padgett, R.W., Phillips, S.J., Voliva, C.F., Comer, M.B., Hardies, S.C., Haigwood, N.L., Langley, C.H., Racine, R.R. and Hutchison, C.A. (1980). The mouse β hemoglobin locus. In: Organization and expression of globin genes. (G. Stamatoyannopoulos and A.W. Nienhuis, Alan R. Liss, Inc., Eds.) pp. 69-88.
- 16 Long, E.O., Rebbert, M.L. and Dawid, I.B. (1981). Proc. Natl. Acad. Sci. U.S.A. 78, 1513-1517.
- 17 Berk, A.J. and Sharp, P.A. (1977). Cell 12, 721-732.
- 18 Kan, N.C. and Gall, J.G. (1982). J. Mol. Biol. in press.
- 19 Jensen, O.E., Paludan, K., Hyldig-Nielsen, J.J., Jorgensen, P. and Marcker, K.A. (1981). Nature 291, 677-679.
- 20 Long, E.O. and Dawid, I.B. (1979). Cell 18, 1185-1196.
- 21 Rae, P.M.M., Barnett, T. and Murtif, V.L. (1981). Chromosoma 82, 637-655.
- 22 Rae, P.M.M. (1981). Nucl. Acids Res. 9, 4997-5010.