3508–3510 Nucleic Acids Research, 1994, Vol. 22, No. 17

A comparative database of group I intron structures

Simon H.Damberger and Robin R.Gutell*
MCD Biology, Campus Box 347, University of Colorado, Boulder, CO 80309-0347, USA

ABSTRACT

We have created a database of comparatively derived group I intron secondary structure diagrams. This collection currently contains a broad sampling of phylogenetically and structurally similar and diverse structures from over 200 publicly available intron sequences. As more group I introns are sequenced and added to the database, we anticipate minor refinements in these secondary structure diagrams. These diagrams are directly accessible by computer as well as from the authors.

INTRODUCTION

Group I introns are a family of RNA molecules that catalyze their own excision. Since it was discovered that group I introns self-splice [1], much experimental and comparative research has been devoted to determining their structure [2]. This work has, in turn, helped us to understand their function during the splicing reaction.

As the number of group I intron sequences has increased, comparative sequence analysis has established and refined the group I intron secondary structure and is now revealing the beginnings of a three dimensional model. [3,4,5,6] To further refine the structural detail of group I introns, we will need continued improvements of our correlation analysis methods. We will also need an even larger and more diverse collection of group I intron sequences.

Currently there are over 200 publicly available group I intron sequences in GenBank [7]. These introns have been found within two of the three primary phylogenetic domains and within the Eucarya nucleus, chloroplast, and mitochondrion. With the current number and expected increase of group I intron sequence availability, the need for a group I secondary structure database has become great.

OBJECTIVES

Our current understanding of the patterns of sequence conservation and variation for the group I intron family of RNA molecules makes it possible to infer their secondary and tertiary structure base pairings. The growing number of sequences presents us with the opportunity to continue to evaluate and refine these structures as additional comparative evidence becomes available. These structures will be provided in a new common format [8] that will allow a larger scientific audience to readily compare similar and diverse introns. The group I secondary structure database has several primary objectives:

1. To collect and structure group I intron sequences as they become available. As more introns are sequenced, we expect to refine the current structures using comparative analysis. Although we expect the core structure of all group I introns to largely remain the same, some minor changes in the structure are possible. The variable regions, however, will probably undergo more revision as sequences continue to become available for comparison. Further comparative analysis will also be called upon to identify new tertiary interactions as comparative algorithms improve and the database grows.
2. To present other information pertaining to group I introns such as a list of all available group I intron sequences, reference lists for intron sequences, cellular location of introns, and their phylogenetic position.
3. To provide these structures to the general community via anonymous ftp (file transfer protocol) and the World Wide Web [9,10]. (Hard copies will also be available to those who do not have access to the Internet.) Initially we will offer representative structures from each of the major subgroups. With time, we will generate more structures to round out the phylogenetic diversity of the database. Within this communication, we also present several examples of introns including the *Ankistrodesmus stipitatus* LSU rRNA intron [11] (fig. 1), *Saccharomyces cerevisae* mitochondrial LSU rRNA intron [12] (fig. 2), and the second *Podospora anserina* mitochondrial apocytochrome b intron [13] (fig. 3).

DESCRIPTION OF DATABASE

The database currently contains 219 intron sequences, the majority of which are available from GenBank. The total number of introns in the different subgroups [as proposed in 6] are listed in Table 1. These introns vary greatly in their phylogenetic diversity and are found in two of the three major phylogenetic domains, Eucarya and (eu)Bacteria [14,15] (this phylogenetic nomenclature is presented in [16]). Within the Eucarya, introns have been found in the nucleus and the two major organelles, the mitochondrion and chloroplast. Group I introns have been identified in 23 different genes (Table 1).

The secondary structures will be available in a new format that more accurately represents the current knowledge of the group I intron three dimensional structure [8]. The secondary structures will be available through three channels; hard copies, anonymous ftp, and through the WWW (World Wide Web). The computer files will only be distributed in the PostScript format, therefore

*To whom correspondence should be addressed

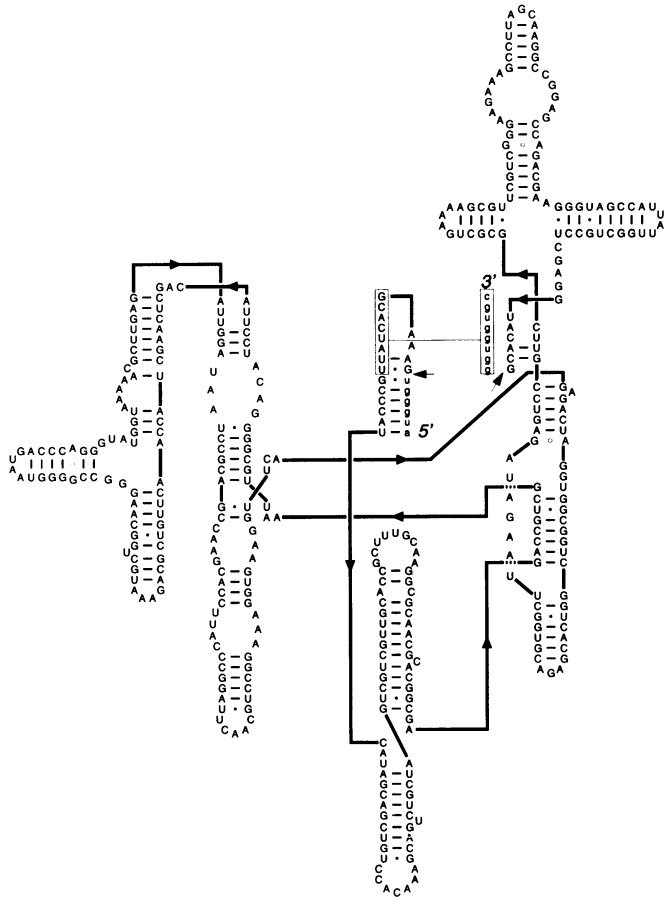

Figure 1. Secondary structure diagram for the SSU rRNA *Ankistrodesmus stipitatus* (intron in capital letters) intron and flanking exon sequences (lower case letters). The arrows point to the 5′ and 3′ splice sites. The thick lines show the continuity of the strands with the arrowheads within the lines showing 5′ to 3′ sequence direction. The boxed nucleotides connected by a thin line represent a base-pairing interaction involved in the exon-ligation step (second step) of RNA splicing [4]. GenBank accession number is X56100.

a printer or viewer that can read PostScript is required to visualize these diagrams. Those users with Internet access will be able to retrieve copies of the structures through ftp or the WWW.

PostScript files of the secondary structure diagrams can be obtained by anonymous ftp at the following site and directory:

ftp address: *pundit.colorado.edu (128.138.212.53)*
directory: */pub/RNA/GRPI/*

The WWW is a system of software and protocols initiated by CERN (European Laboratory for Particle Physics) to facilitate the access of information via the Internet. It uses a hypertext and multimedia presentation to simplify the accessibility of network data. In order to use the WWW, one must first retrieve a program that navigates the WWW like the NCSA (National Center for Supercomputing Applications) Mosaic WWW browser.

NCSA Mosaic is a user-friendly WWW browser that is available for Microsoft Windows, Macintoshes, and UNIX workstations that use the X-Windows interface. NCSA Mosaic can be retrieved by anonymous ftp from the ftp site *ftp.ncsa.uiuc.edu (141.142.20.50)* in the */Mosaic* directory.

To obtain PostScript files of the secondary structures through the WWW, please use the following URL:

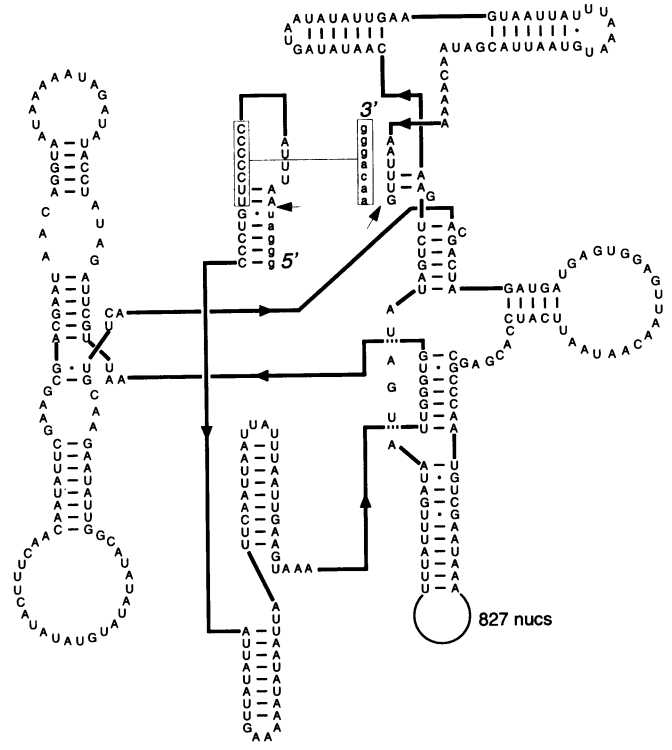

Figure 2. Secondary structure diagram for the LSU rRNA *Saccharomyces cerevisiae* mitochondrial intron. GenBank accession number is X00149.

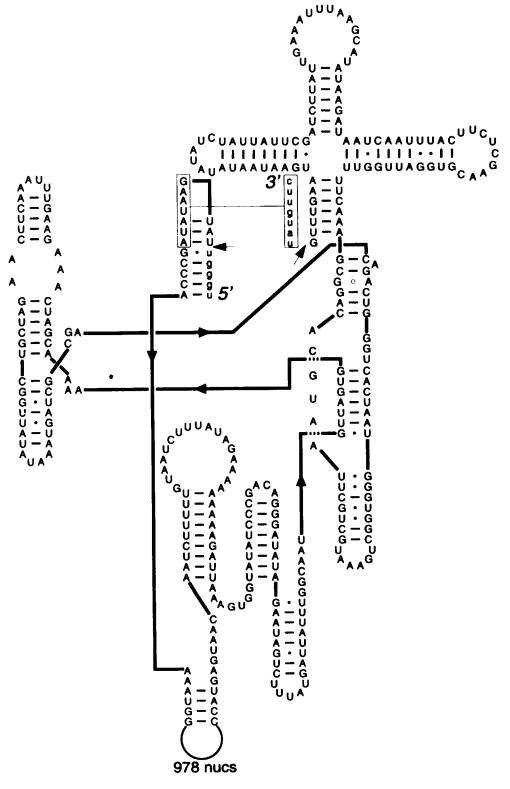

Figure 3. Secondary structure diagram for the second apocytochrome b *Podospora anserina* mitochondrial intron. GenBank accession number is X55026.

Table 1.

Subgroup	Number	Genes represented
IA1	28	ATP9; CYTB; LSU rRNA; ND1; OX2; psbA; psbB; psbC
IA2	6	g31; nrdB; nrdD; psbC; td
IA3	7	LSU rRNA; SSU rRNA
IB1	11	CYTB; LSU rRNA; OX1; SSU rRNA
IB2	17	ATP6; LSU rRNA; ND1; ND5; OX1
IB3	13	LSU rRNA; OX1; SSU rRNA
IB4	13	CYTB; LSU rRNA; ND5; OX1; psbA
IC1	46	LSU rRNA; ND1; ND4L; SSU rRNA
IC2	8	ATP6; ND1; ND3; ND4; ND5; OX1; SSU rRNA
IC3	33	LSU rRNA; OX1; tRNA arg; tRNA ile; tRNA leu
ID	13	CYTB; OX1; OX3; ND5
Unknown	24	CYTB; ND1; ND5; OX1; OX2; psbA; SSU rRNA;
Total	219	

URL: *http://pundit.colorado.edu:8080/RNA/GRPI/introns.html*

The WWW database is formatted to make it easy for the user to find the intron or introns of interest. There are two methods of accessing the secondary structure diagrams. The first method is through links set up by subgroup and phylogeny where the user can click on the subgroup of interest and look for introns through a list classified by phylogeny. For the second method, the database allows boolean 'and' searches over all introns in the database by organism name, intron site (gene and cellular location), and subgroup (see WWW examples). These facilities will be updated and enhanced as more introns are sequenced and structured, and as our database expands. We welcome any suggestions on how to improve the search facilities as well as the database itself.

WWW examples

1. To find all subgroup IA1 introns that occur in *Podospora anserina*, enter the following into the search box (all searches are case insensitive):
   Podospora anserina IA1
2. To find all mitochondrial LSU rRNA introns, enter the following:
   LSUrRNA Mitochondrion

This publication should be quoted as a reference for any secondary structure data obtained either through hard copies from the authors or from the database. Requests for hard copy printouts should be referred to RRG. Questions and comments about the on-line resources should be directed to SHD.

ACKNOWLEDGEMENTS

We would like to thank Tom Cech for encouraging us to pursue this project. We acknowledge Bryn Wiser and Tom Macke for access to the programs that facilitate the generation of the structures (XRNA) and alignments (AE2). We would also like to thank the W. M. Keck Foundation for their generous support of RNA science on the Boulder campus and Sun Microsystems for the donation of computer equipment. This work was supported by grants from the NIH (GM48207) and the Colorado RNA Center to R. R. G.

REFERENCES

1. Cech, T. R., Zaug, A. J., and Grabowski, P. J. *In vitro* splicing of the ribosomal RNA precursor of *Tetrahymena*: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* **27**, 487–496 (1981).
2. Cech, T. R. Self-splicing of group I introns. *Annual Reviews of Biochemistry* **59**, 543–568 (1990).
3. Michel, F., Jacquier, A., and Dujon, B. Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie* **64**, 868–881 (1982).
4. Davies, R. W., Waring, R. B., Ray, J. A., Brown, T. A., and Scazzocchio, C. Making ends meet: A model for RNA splicing in fungal mitochondria. *Nature* **300**, 719–724 (1982).
5. Cech, T. R. Conserved sequences and structures of group I introns: building an active site for RNA catalysis — a review. *Gene* **73**, 259–271 (1988).
6. Michel, F., and Westhof, E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *Journal of Molecular Biology* **216**, 585–610 (1990).
7. Benson, D., Lipman, D. J., and Ostell, J. GenBank. *Nucleic Acids Research* **21**, 2963–2965 (1993).
8. Cech, T. R., Damberger, S. H., and Gutell, R. R. Representation of the secondary and tertiary structure of group I introns. *Nature Structural Biology* **1**, 273–280 (1994).
9. Berners-Lee, T. J., Cailliau, R., and Groff, J. -F. The World-Wide Web, *Computer Networks and ISDN Systems* **25**, 454–459 (1992).
10. Krol, E. *The Whole Internet Catalog and User's Guide*, O'Reilly and Associates, Inc. pp. 227–242 (1992).
11. Davila-Aponte, J. A., Huss V. A. R., Sogin, M. L., and Cech, T. R. A self-splicing group I intron in the nuclear pre-rRNA of the green alga, *Ankistrodesmus stipitatus*. *Nucleic Acids Research* **19**, 4429–4436 (1991).
12. Jacquier, A., and Dujon, B. The intron of the mitochondrial 21S rRNA gene: Distribution in different yeast species and sequence comparison between *Kluyveromyces thermotolerans* and *Saccharomyces cerevisiae*. *Molecular General Genetics* **192**, 487–499 (1983).
13. Cummings, D. J., Michel, F., McNally, K. L. DNA sequence analysis of the apocyochrome b gene of *Podospora anserina*: a new family of intronic open reading frame. *Current Genetics* **16**, 407–418 (1989).
14. Xu, M. -Q., Kathe, S. D., Goodrich-Blair, H., Nierzwicki-Bauer, S. A., and Shub, D. A. Bacterial origin of a chloroplast intron: conserved self-splicing group I introns in cyanobacteria. *Science* **250**, 1566–1570 (1990).
15. Kuhsel, M. G., Strickland, R., and Palmer, J. D. An ancient group I intron shared by eubacteria and chloroplasts. *Science* **250**, 1570–1572 (1990).
16. Woese, C. R., Kandler, O., and Wheelis, M. L. Toward a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences USA* **87**, 4576–4579 (1990).