Protein Science (1994), *3*:1117–1120. Cambridge University Press. Printed in the USA.

FOR THE RECORD

Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family

ALEXANDER E. GORBALENYA

Institute of Poliomyelitis and Viral Encephalitides, Russian Academy of Medical Sciences, 142782 Moscow Region, Russia, and Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907-1392

(RECEIVED February 28, 1994; ACCEPTED May 4, 1994)

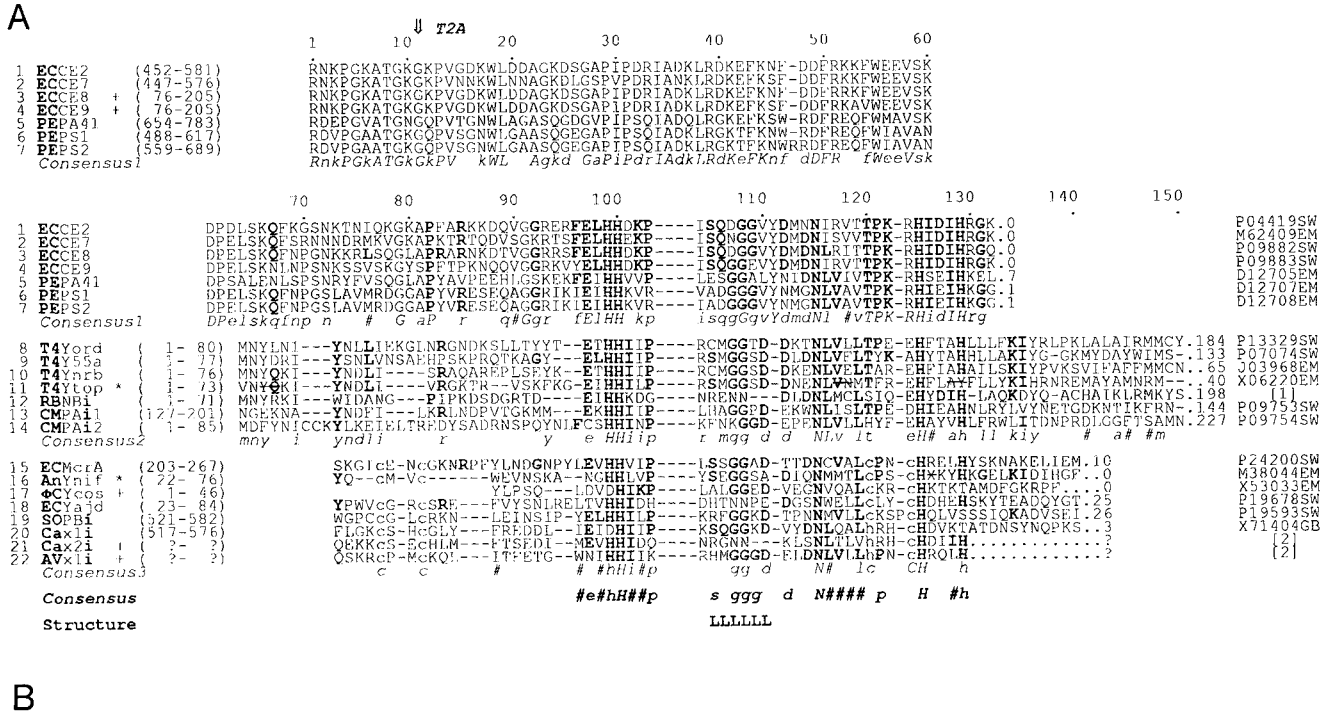
Abstract: A new family of protein domains consisting of 50–80 amino acid residues is described. It is composed of nearly 40 members, including domains encoded by plastid and phage group I introns; mitochondrial, plastid, and bacterial group II introns; eubacterial genomes and plasmids; and phages. The name "EX_1HH-HX_3H" was coined for both domain and family. It is based on 2 most prominent amino acid sequence motifs, each encompassing a pair of highly conserved histidine residues in a specific arrangement: EX_1HH and HX_3H. The "His" motifs often alternate with amino- and carboxy-terminal motifs of a new type of Zn-finger-like structure $CX_{2,4}CX_{29\text{-}54}[CH]X_{2,3}[CH]$. The EX_1HH-HX_3H domain in eubacterial E2-type bacteriocins and in phage RB3 (wild variant of phage T4) product of the nrdB group I intron was reported to be essential for DNA endonuclease activity of these proteins. In other proteins, the EX_1HH-HX_3H domain is hypothesized to possess DNase activity as well. Presumably, this activity promotes movement (rearrangement) of group I and group II introns encoding the EX_1HH-HX_3H domain and other gene targets. In the case of *Escherichia coli* restrictase McrA and possibly several related proteins, it appears to mediate the restriction of alien DNA molecules.

Keywords: E2-type bacteriocins; mcrA restriction; *nifD* locus rearrangement; phage ΦC31; phage T4 ORFs; zinc finger

Self-splicing introns containing open reading frames (ORFs) are suggested to represent a sort of molecular commensalist that probably emerged relatively recently in the course of evolution. In the partnership, introns provide a convenient genetic system for maintaining and expression of protein-encoded sequences, and the products of ORFs assist introns in their splicing and/or promote intron movement in the genetic environment. The perfect and very puzzling correlation between the group of self-splicing introns and the family of proteins "infecting" it has been documented (Lambowitz & Belfort, 1993). Thus, group I introns encode proteins of several families exhibiting site-specific DNA endonuclease (DNase) activity important for intron movement. In contrast, group II encoded polypeptides contain reverse transcriptase (RT), X, and Zn-finger-like (Zn) domains (Mohr et al., 1993). Intronic RT activity is believed to mediate group II intron transfer.

I report here a case of the unprecedented relationship between proteins encoded by introns of the two groups. Initially, through pairwise comparisons, a small region common for 2 *Chlamydomonas moewusii* plastid group I intronic ORFs **CM**PAi1 and **CM**PAi2 (see Fig. 1 caption for explanation of sequence nomenclature) and a phage RB3 group I intronic endonuclease **RB**NBi was delineated (Fig. 1A, entries 12–14). Subsequently, these proteins were compared with the sequence databanks in an iterative manner, which allowed a total of nearly 20 sequences containing the same conserved domain recognized earlier to be retrieved (Fig. 1A). The sequences were arbitrarily sorted among 3 groups according to pairwise sequence similarity level and/or unique common features. The first group consists of closely related (probability [*P*] of matching by chance $< 10^{-16}$) DNase domains of bacteriocins of the E2-type and was described earlier (Sano et al., 1993). The second group comprises 3 group I intronic sequences and 4 phage T4 ORFs; in an overwhelming majority of the cases, a similarity between the latter and **CM**PAi1 was statistically significant ($P < 10^{-3}$), whereas **CM**PAi2 and **RB**NBi did not show clear sequence affinities. The third, most divergent group of sequences with a limited number of statistically significant matches was delineated based on the presence of 2 Zn-finger-like motifs, CX_2C and $[CH]X_{2,3}C$. It includes 2 *Escherichia coli* gene products, McrA restrictase and ORF6 in SecD locus (**EC**Yajd), *Anabaena* unidentified reading frame (URF) associated with the NifD locus, 4 Zn domains of group II intron-encoded RT-like sequences of bacterial and plastid origin, and phage ΦC31 URF. A similarity between McrA restrictase and the Zn domains was recognized earlier (Ferat & Michel, 1993). Unlike other sequences, **EC**Yajd has been retrieved from the SWISS-PROT database with the help of a profile constructed from sequences of the second and third groups.

Reprint requests to: Alexander E. Gorbalenya, Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907-1392; e-mail: aeg@bragg.bio.purdue.edu.



B

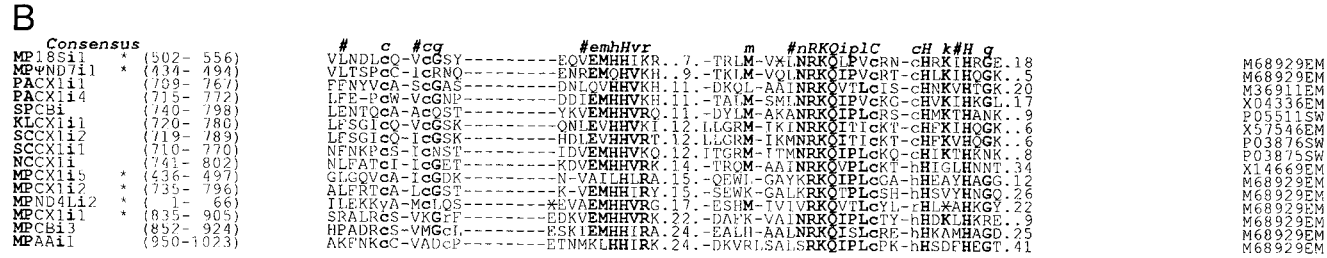


Fig. 1. Amino acid sequence alignments of the $EX_1HH\text{-}HX_3H$ domains. Pairwise local and multiple global sequence alignments were generated with the help of the DotHelix (Leontovich et al., 1993) and OPTAL (Gorbalenya et al., 1989) programs, respectively. The sequence databases were searched utilizing the GCG package (Devereux et al., 1984), the BLITZ program (Sturrock & Collins, 1993), a family of the BLAST programs (Altschul et al., 1990), and the QUICK program in the GeneBee shell (Brodsky et al., 1993). Secondary structure predictions were done with the help of the PHD program (Rost & Sander, 1993). Default values were used for all parameters. The positions of the aligned domains in the polypeptides are indicated, and the distance to the carboxy-terminal end of a protein is shown. For the entries designated by a plus sign, a complete sequence is not yet available. In the sequences designated by an asterisk, readthrough (*rt*) translation of termination codon(s) and/or frameshifting (*fs*) within the region corresponding to the domain and/or upstream of it were used to maintain an ORF. Two residues, between which the sequence was assumed to be frameshifted, and those of the plausible *rt* stop-codons designated by X are depicted in the strikethrough style. The numbering for some of these sequences starts from the first residue of the composite ORF rather than from a putative initiator Met residue. The Cys/His residues conforming to a Zn-finger-like structure and its apparent 1-nucleotide derivatives, Arg and Tyr, are shown in lower case. The consensus line depicts invariant residues in upper case, and those which, alone or in a group, are present in not less than half of the sequences are shown in lower case or are designated by a number symbol (#) used for hydrophobic residues (I, L, V, M, F). The structure line shows secondary structure prediction: L, loop. Sequence nomenclature: The first 2 characters designate species; the next 2–4 characters depict gene (ORF) name; the next optional character **i** is shown in bold, which is sometimes followed by a number, and collectively denotes self-splicing intron. For the sake of uniformity, phage RB3 I-TevIII (intron-encoded T-even endonuclease III) was renamed **RB**NBi. *Abbreviations of species:* Eukaryotes: CM, *Chlamydomonas moewusii*; KL, *Kluyveromyces lactis*; MP, *Marchantia polymorpha*; NC, *Neurospora crassa*; PA, *Podospora anserina*; SC, *Saccharomyces cerevisiae*; SO, *Scenedesmus obliquus* (originally Green alga KS3/2); SP, *Schizosaccharomyces pombe*. Eubacteria: An, *Anabaena* sp. strain PCC7120; AV, *Azotobacter vinelandii* UWR; Ca, *Calothrix* PCC7601; EC, *Escherichia coli*; PE, *Pseudomonas aeruginosa*. Phages: RB, *E. coli* phage RB3, wild isolate of T4; T4, *E. coli* phage T4; ΦC, *Streptomyces* phage ΦC31. *Abbreviations of genes/ORFs/URFs:* Mitochondrial: 18S, 18S rRNA; AA, ATPase F1 subunit α; CB, apocytochrome *b*; CX1, cytochrome oxidase subunit 1; ND4L, putative subunit of NADH ubiquinone reductase; ΨND7, pseudogene for putative subunit of NADH ubiquinone reductase. Chloroplast: PA, thylakoid membrane protein D1 of photosystem II; PB, subunit IV of the cytochrome b_6/f complex. Eubacterial: CE2–CE9, plasmid-encoded colicin E2–E9, respectively; mcrA, $C^{me}CGG$ restrictase; PA41, PS1, and PS2, pyocin A41, S1, and S2, respectively; X1 and X2, ORFs of an undefined function and uncertain localization; Yajd, ORF6 in *secD* operon for integral protein of the cytoplasmic membrane; Ynif, URF in the 11-kbp DNA element of *nifD* for α-subunit of nitrogenase. Phage: NB, *nrdB* for chain B2 of the ribonucleotide reductase; Y55a, ORF 55.10 overlapping the terminator codon of *sunY* for the presumed anaerobic ribonucleotide reductase; Ycos, URF adjacent to COS region; Ynrb, ORF nrdB.2 overlapping Ter codon of *nrdA* for chain B1

(*continued on facing page*)

Only a few intergroup relationships were statistically significant, including pairs **PE**PA41 and **EC**McrA, and **AV**x1i and **T4**Y55a. However, within an approximately 35-amino acid region the patterns of conserved residues of all three groups are certainly similar, demonstrating that all sequences share a common domain. The finding of the Zn domain among the retrieved proteins prompted an inspection of the rest of the group II intronic domains of this kind, all of which were shown to be encoded by mitochondrial (mt) genomes. Two conserved motifs enriched in histidine residues, EX_1HH (amino-terminal) and HX_3H (carboxy-terminal), were recognized both in mt and nonmitochondrial (non-mt) sequences, and 2 Zn-finger-like motifs, $CX_{2,4}C$ and $[CH]X_{2,3}[CH]$, in the third group of proteins and mt sequences (Fig. 1). A fraction of mt sequences contains replacement(s) of otherwise conserved His and/or Cys residues of the Zn-finger-like motifs. In those sequences, *rt* stop-codons and/or *fs* were used to maintain an ORF, implying that sequencing mistakes or some type of (post)transcriptional modifications could take place or that those sequences are remnants rather than active proteins. Mitochondrial sequences strikingly differ from other proteins in a region between the EX_1HH and $[CH]X_{2,3}[CH]$ motifs. In light of this fact, a relationship between mt and non-mt sequences might be apparent. However, this is unlikely because the Zn domain occupies the same most C-terminal position in the RT-containing intronic proteins of mt and non-mt origin, and 4 conserved motifs are organized in similar linear order in all proteins of this sort (Fig. 1). Thus, the observed dissimilarity could be more likely a result of extensive divergent evolution from a common ancestor, and the mt Zn domain can be treated as a variant of the domain delineated above. This domain will be referred to as the EX_1HH-HX_3H domain, after the 2 most conserved motifs it contains.

The EX_1HH-HX_3H domain is likely a conserved core of DNase, since this is a large portion of bacteriocin DNases (Schaller & Nomura, 1976), and a part of a fragment of the **RB**NBi, which is indispensable for enzymatic activity of the intronic DNase (Eddy & Gold, 1991). DNases of the new family can or could mediate a number of processes, including restricting alien DNA (McrA restrictase, bacteriocins, phage T4 ORFs) and conferring mobility to genetic elements (intronic sequences, nifD locus). For the first time, (putative) DNases were tentatively identified among group II intron-encoded proteins and found to be similar to those encoded by group I introns. This implies that mechanisms of group I and group II intron transfer may have more in common than suspected earlier. Particularly, the intron movement may be initiated by intron-encoded DNase through a double-strand break not only in the case of group I introns but in group II introns as well.

The EX_1HH and HX_3H motifs probably form the catalytic center of DNase and include catalytic residue(s) that are most likely 1 or 2 invariant histidine residues. In a large fraction of the proteins, EX_1HH and HX_3H motifs alternate with $CX_{2,4}C$ and $[CH]X_{2,3}[CH]$ motifs. Such overlapping of fingerlike and putative enzymatic domains has not been encountered previously in Zn-finger proteins. Whatever its type, the fold adopted by the EX_1HH-HX_3H domain is believed to be essentially similar in proteins with and without Zn-binding potential. This probably indicates that the spatial structure of a finger domain in the region connecting $CX_{2,4}C$ and $[CH]X_{2,3}[CH]$ motifs can also be maintained by residues other than those that bind a Zn cation.

In conclusion, the third largest family of DNases encoded by self-splicing introns was described. In many respects, the EX_1HH-HX_3H family resembles the LAGLI-DADG and GIY-YVG families, which cover group I intronic and nonintronic proteins (Hensgens et al., 1983; Michel & Dujon, 1986). However, in sharp contrast with previous experience, the new family unites proteins encoded by both the group I and group II introns. Additionally, a tentative assignment of DNase activity to the Zn domain of group II introns strengthens the parallel in domain organization between intronic and retroviral RT-containing polypeptides (Doolittle et al., 1989; Mohr et al., 1993). This indicates that intronic Zn and retroviral integrase domains may be functionally similar.

Acknowledgments

My thanks are due to Leonid Brodsky for assistance with the database searches in the GeneBee shell; Konstantin Chumakov, Eugene Koonin, and Mikhail Rozanov for making available to me primary sources of information used throughout the work; Eric Snijder for critical reading of a draft of the paper; Vadim Agol, Michael Rossmann, and Willy Spaan for encouragement; and Sharon Wilder for help with the preparation of the manuscript. My stay at Purdue University was supported

Fig. 1. *Continued.*

of the ribonucleotide reductase; Yord, ORF D (αgt.1) between *αgt* for glucosyltransferase and *55* for protein necessary for late RNA synthesis; Ytop, composite ORF overlapping 3′-end of *topB* for large chain of the type II topoisomerase. All but 3 sequences were from EMBL (EM), SWISS-PROT (SW), or GENBANK (GB) databases, and the corresponding accession numbers are shown rightmost; 1, Eddy and Gold (1991); 2, Ferat and Michel (1993). Three gaps of variable length were arbitrarily introduced in mt sequences to align Zn-finger and "His"-rich motifs of non-mt and mt proteins, and the length for one of these gaps is indicated. **A:** Amino acid sequence alignments of the EX_1HH-HX_3H domains encoded by non-mt genomes. Throughout the multiple-stage aligning procedure, each alignment produced was scored over 20 standard deviations (SD) (group 1, entries 1–7), 6.2 SD (group 2, entries 8–14; the C-terminal 19 amino acids portion of the alignment is not shown), and 5.3 SD (group 3, entries 15–19) compared with an average score for scrambled sequences. Entries 20–22 were appended to those of 15–19 following the alignment of Ferat and Michel (1993). The position of a gap in the Gly-rich region of the alignments of the **RB**NBi and **Ca**x2i sequences was slightly changed to underscore their similarity. The residues found in more than half of the sequences belonging to one group and, additionally, in at least 1 sequence in the other groups are highlighted in bold. The cumulative consensus for all 3 groups depicts invariant residues in italic bold upper case. Those found in more than half of the sequences of at least one group and additionally present in sequences of the other groups are shown in italic bold lower case. Secondary structure predictions are shown for the region where the most consistent results were obtained with different sequences. The N-terminal border of the T2A proteolytic fragment of colicin E2 molecules possessing DNase activity is marked by an arrowhead. **B:** Amino acid sequence alignment of the EX_1HH-HX_3H domains encoded by mitochondrial genomes. Throughout alignment, scores were obtained that were not below 8.9 SD. The residues found in more than half of the sequences are highlighted in bold.

by a Cooperation in Applied and Science Technology award to A.E.G. and Michael Rossmann, and an NSF grant to Michael Rossmann. I apologize to all researchers whose relevant works were not properly cited due to space limitations.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403-410.

Brodsky LI, Drachev AL, Leontovich AM, Feranchuk SI. 1993. GeneBee: The program package for biopolymer structure analysis. In: Gindikin S, ed. *Mathematical methods of analysis of biopolymer sequences*. Providence, Rhode Island: American Mathematical Society. pp 127-140.

Devereux J, Haeberli P, Smithies O. 1984. A comprehensive set of sequence programs for the VAX. *Nucleic Acids Res 12*:387-395.

Doolittle RF, Feng DF, Johnson MS, McClure MA. 1989. Origins and evolutionary relationships of retroviruses. *Q Rev Biol 64*:1-30.

Eddy SR, Gold L. 1991. The phage T4 nrdB intron: A deletion mutant of a version found in the wild. *Genes Dev 5*:1032-1041.

Ferat JL, Michel F. 1993. Group II self-splicing introns in bacteria. *Nature (Lond) 364*:358-361.

Gorbalenya AE, Blinov VM, Donchenko AP, Koonin EV. 1989. An NTP-binding motif is the most conserved sequence in a highly diverged monophyletic group of proteins involved in positive strand RNA viral replication. *J Mol Evol 28*:256-268.

Hensgens LAM, Bonen L, de Haan M, van der Horst G, Grivell LA. 1983. Two intron sequences in yeast mitochondrial COX1 gene: Homology among URF-containing introns and strain-dependent variation in flanking exons. *Cell 32*:379-389.

Lambowitz AM, Belfort M. 1993. Introns as mobile genetic elements. *Annu Rev Biochem 62*:587-622.

Leontovich AM, Brodsky LI, Gorbalenya AE. 1993. Construction of the full local similarity map for two biopolymers. *BioSystems 30*:57-63.

Michel F, Dujon B. 1986. Genetic exchanges between bacteriophage T4 and filamentous fungi? *Cell 46*:323.

Mohr G, Perlman PS, Lambowitz AM. 1993. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res 21*:4991-4997.

Rost B, Sander C. 1993. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA 90*:7558-7562.

Sano Y, Matsui H, Kobayashi M, Kageyama M. 1993. Molecular structures and functions of pyocins S1 and S2 in *Pseudomonas aeruginosa*. *J Bacteriol 175*:2907-2916.

Schaller K, Nomura M. 1976. Colicin E2 is a DNA endonuclease. *Proc Natl Acad Sci USA 73*:3989-3993.

Sturrock SS, Collins JF. 1993. *MPsrch version 1.3*. Edinburgh, UK: Biocomputing Research Unit, University of Edinburgh.