Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 6798–6802, August 1992
Biochemistry

Presence of a member of the Tc*1*-like transposon family from nematodes and *Drosophila* within the vasotocin gene of a primitive vertebrate, the Pacific hagfish *Eptatretus stouti*

(exon–intron organization/neurophysin/copeptin/cyclostome)

JÖRG HEIERHORST*, KARL LEDERIS†, AND DIETMAR RICHTER*‡

*Institut für Zellbiochemie und klinische Neurobiologie, Universität Hamburg, UKE, Martinistrasse 52, 2000 Hamburg 20, Federal Republic of Germany; and †Department of Pharmacology and Therapeutics, 2500 University Drive Northwest, Calgary, Alberta, Canada T2N 1N4

Communicated by Hans J. Müller-Eberhard, March 25, 1992 (received for review February 12, 1992)

ABSTRACT **Molecular cloning of the vasotocin gene of a cyclostome, the Pacific hagfish *Eptatretus stouti*, reveals, in contrast to other known members of the vertebrate vasopressin/oxytocin hormone gene family, an unusual exon–intron organization. Although the location of three exons and two introns is conserved, an additional intron is present 5′ of the coding region of the hagfish gene. The third intron, which is >14 kilobase pairs in size, contains on the opposite DNA strand to that encoding vasotocin an open reading frame exhibiting striking similarity to the putative transposase of Tc*1*-like nonretroviral mobile genetic DNA elements, so far reported only from nematodes and *Drosophila*. The hagfish element, called Tes*1*, is flanked by inverted terminal repeats representing an example of the existence of a typical inverted terminal-repeat transposon within vertebrates. The presence of Tc*1*-like elements in nematodes, *Drosophila*, and cyclostomes indicates that these genetic elements have a much broader phylogenetic distribution than hitherto expected.**

The jawless cyclostomes—hagfish and lampreys—are descendants of the most primitive living vertebrates, which diverged from the mainstream of vertebrate phylogeny >500 million years ago (1). Of all vertebrates, they have the simplest central nervous systems (2), making them attractive model systems for studying various aspects of brain function (3). The cyclostome vasotocin gene structure is of evolutionary interest because the nonapeptide vasotocin is supposed to represent the common ancestor of the vasopressin/oxytocin superfamily found in higher vertebrates (4). The coexistence of at least one peptide hormone from each peptide family (vasopressin-like and oxytocin-like) in all higher vertebrate classes is, thus, thought to result from duplication and consecutive divergence of a common ancestral gene before the radiation of cartilaginous fish, which occurred some 450 million years ago (4).

Classification of the nonapeptides that have the consensus sequence Cys-Xaa-Xaa-Xaa-Asn-Cys-Pro-Xaa-Gly-amide as a member of one or other of these families, is based solely on structural considerations; although a basic amino acid residue at position 8 is typical of vasopressin family members, this residue is substituted by a hydrophobic residue in oxytocin-like hormones. In mammals, vasopressin has a key endocrine role in hydroosmotic regulation, whereas oxytocin serves a reproductive function. Nothing is known about the endogenous function of the hagfish vasotocin-like peptide; however, an osmohomeotic function can almost certainly be excluded because the plasma of these animals is known to be isosmolar with the surrounding seawater (5).

Previous studies have shown that the precursor molecules of both hormone families consist of a signal peptide, the hormone sequence, a Gly-Lys-Arg processing and modification sequence, and a carrier molecule termed neurophysin (6–9). All precursors of the vasopressin family are extended by a so-called copeptin domain, which is also present in the sucker isotocin precursors (8); the latter are the equivalents of mammalian oxytocin. In mammals and amphibia, this domain has been deleted from the oxytocin precursor. The common evolutionary origin of the two peptide families is further emphasized by their similar gene structures. Mammalian vasopressin and oxytocin genes each contain two introns, which interrupt the coding regions for the neurophysins at similar positions (6).

Based on cDNA and gene sequence data from the Pacific hagfish, this report provides conclusive evidence for the existence of the vasotocin precursor[§] in this species and describes the organization of this gene, which, in contrast to those of higher vertebrates, contains an additional intron that interrupts the 5′-untranslated region. Remarkably, an open reading frame (ORF) has been detected within the third intron that predicts a protein strikingly similar to the putative transposases of Tc*1*-like elements normally found in nematodes and *Drosophila* (10–14).

MATERIALS AND METHODS

RNA and Genomic DNA Preparations. Poly(A)$^+$ RNA was prepared from either whole brains or dissected mesencephali of several Pacific hagfish (*Eptatretus stouti*) by using standard procedures (15, 16). Genomic DNA was prepared from individual hagfish livers (17).

Labeling of Brain Poly(A)$^+$ RNA. One microgram of poly(A)$^+$ RNA from hagfish whole brains was transcribed (17) into 100 ng of ^{32}P-labeled cDNA probe by using [α-^{32}P]dCTP and random hexanucleotide primers (Boehringer Mannheim). After alkali hydrolysis of the template RNA, the probe (5×10^8 cpm/μg of cDNA) was used at a concentration of 20 ng of cDNA per ml of hybridization solution.

cDNA Library Construction, Screening, and Sequence Analysis. Five micrograms of hagfish hypothalamic poly(A)$^+$ RNA was used for cDNA synthesis (18), which was subsequently blunt-ended by using T4 DNA polymerase, ligated to *Eco*RI adaptors, and inserted into the *Eco*RI site of the λZAP II cloning vector (19), yielding 4×10^5 recombinants. After amplification, the DNA from 10^5 recombinants was transferred onto nitrocellulose membranes and hybridized, in duplicate, with 50 pmol of a 96-fold degenerate 20-base

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: ORF, open reading frame.

‡To whom reprint requests should be addressed.

§The sequences reported in this paper have been deposited in the GenBank data base (accession nos. M93037–M93040).

oligonucleotide (5′-GGRCARTTYTGDATRTARCA-3′, in which R = A or G; N = A, C, G, or T; D = A, G, or T; and Y = C or T), which is complementary to all possible codons for the first seven amino-acid residues of vasotocin. The oligonucleotide was end-labeled to a specific activity of $>10^6$ cpm/pmol. Replica filters were washed finally in 3 M $(CH_3)_4NCl$ at 58°C (20). A single vasotocin-encoding clone was isolated and upon sequence analysis appeared to lack the entire 5′-untranslated sequence and the first part of the putative signal peptide. This cDNA clone labeled to a specific activity of 10^9 cpm/μg (21) was used to screen another 4×10^5 recombinants. Seven out of 30 positively hybridizing clones were purified, and those that gave rise to the longest *Sma* I/*Eco*RI restriction fragments at their 5′ ends were subsequently sequenced. Inserts were rescued as pBluescript(−) phagemids as described (19).

Restriction fragments were subcloned into M13mp18 (22) or pBluescript SK(+) vectors (Stratagene) and sequenced (23) by using the Sequenase kit (United States Biochemical). Genomic DNA sequences were determined by using an automated DNA analysis machine (Applied Biosystems model 304).

Cloning of the Vasotocin Gene. Products [15–24 kilobase pairs (kbp)] of a partial *Sau*3AI digestion of genomic DNA from a single hagfish were cloned into the *Bam*HI sites of the λDASHII vector. Independent clones (2×10^5) were used for infection of the *Escherichia coli* strain SRB, and the resultant plaques were screened with a radiolabeled (21) vasotocin cDNA. A cassette containing the DNA insert flanked by T3 and T7 RNA polymerase promoters was excised from isolated clones with *Not* I. A restriction map of the hagfish vasotocin gene was determined by partial digestion of cloned DNA with various restriction endonucleases and, successively, by hybridizations with T3 and T7 oligonucleotides in Southern blot experiments (24). The locations of exonic sequences were determined by Southern blot analyses of completely digested DNA from genomic clones with cDNA or oligonucleotide probes and with sequence analysis after subcloning into pBluescript SK(+).

RNA and Southern Blot Analysis. Five micrograms of restriction endonuclease-digested genomic DNA or 10 μg of glyoxylated poly(A)$^+$ RNA was separated electrophoretically in 0.7% (wt/vol) or 1.3% (wt/vol) agarose gels and transferred onto Hybond-N membranes (Amersham), according to the supplier's recommendations. Membranes were hybridized with radiolabeled cDNA probes by using standard conditions (17). The final wash stringency was 15 mM sodium chloride/1.5 mM sodium citrate/0.1% (wt/vol) SDS at 65°C. For Southern blot analyses of genomic clones, 1 μg of appropriately digested DNA was loaded per lane.

RESULTS AND DISCUSSION

Structure of the Vasotocin cDNA and the Encoded Precursor. The longest vasotocin cDNA insert consists of 1049 nucleotides, excluding the poly(A) tail (Fig. 1). The first ATG codon is found at positions 52–54, which is followed by an ORF that predicts an authentic vasotocin hormone precursor (Fig. 2) and, remarkably, ends with a cluster of 11 contiguous stop codons at positions 535–567. The 3′-untranslated region comprises 515 base pairs (bp) and has a polyadenylylation signal (AATAAA) located at positions 1026–1031. A transcript of 1150 bases was detected in RNA blot experiments of mesencephalon poly(A)$^+$ RNA (Fig. 3*A*), suggesting that the longest cDNA essentially represents a full-length clone.

The single ORF of the hagfish cDNA predicts a precursor molecule consisting of 161 amino acid residues with a calculated molecular mass of 15,815 Da. The precursor is comprised of a hydrophobic signal peptide of 22 residues followed by the vasotocin hormone moiety, the amino acid motif Gly-Lys-Arg, and a neurophysin moiety that is extended by a copeptin-like sequence. If the lysine residue at position 129 is considered the border between the neurophysin and the putative copeptin, the former should comprise 95 amino acids, the latter 32 residues including the characteristic hydrophobic/basic amino acid core, which is found in related sequences from higher vertebrates (Fig. 2). Unlike all teleost fish copeptin domains examined to date (8, 9), the respective hagfish domain contains a potential N-linked glycosylation signal (residues 143–145). Its location, however, is much closer to the hydrophobic/basic core than glycosylation sites found in mammalian copeptins.

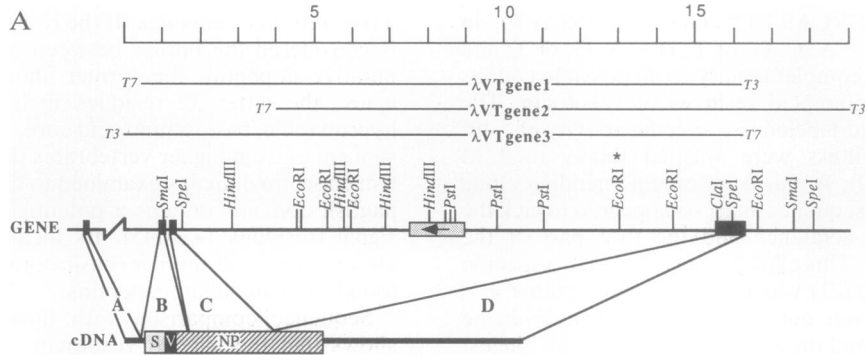
Sequence comparison with those of other vertebrates shows that the hagfish vasotocin prohormone contains 14 conserved cysteine residues within the neurophysin sequence at identical positions, implying that all known neurophysins are folded in a similar manner by specific disulfide bridges. This bonding has been shown to be crucial for establishing the conformation of the bovine neurophysin (25, 26) and, hence, for the latter to correctly bind the peptide hormone (27). Interestingly, a glutamic acid residue at position 47 of the bovine vasopressin-associated neurophysin (indicated by arrow in Fig. 2), whose γ-carboxyl group interacts noncovalently with bound hormone (26), is also conserved in the hagfish neurophysin moiety.

Vasotocin Gene Structure. To analyze the structure of the vasotocin gene a representative genomic liver DNA library was constructed from a single hagfish. Independent clones (2×10^5) from this library were screened with a radiolabeled hagfish vasotocin cDNA probe. Three genomic clones were chosen for detailed restriction and sequence analysis. An *Eco*RI restriction fragment length polymorphism detected among the genomic clones (depicted by a star in Fig. 1*A*) indicates that both alleles of the analyzed individual have been cloned.

Comparison of the cDNA and gene sequences (Fig. 1*B*) reveals that the hagfish vasotocin gene contains two introns at similar positions to those in the mammalian vasopressin and oxytocin genes (6). However, the 5′-most 45 bp of the cDNA were not found within the 862 bp of the genomic sequence preceding that which overlaps with the rest of the cDNA. This result suggests that the hagfish vasotocin gene contains an additional intron compared with related genes in higher vertebrates, 5′ of the coding region. In Fig. 1, the two sequences are aligned so that the end of the putative first exon (A) and the intron 1–exon B junction would correspond to consensus sequences for exon–intron boundaries (28). Thus, this intron, the exact size of which remains to be determined, should be located between nucleotides 45 and 46 of the cDNA interrupting the 5′-untranslated region. Whether this intron evolved *de novo* after the evolutionary radiation of cyclostomes or whether it has been deleted in the line of descent of tetrapods is presently unclear. Due to the lack of teleost vasotocin cDNA clones with long extensions at their 5′ ends (29), the existence of such an intron would not have been detected in the vasotocin gene of *Catostomus commersoni*.

Although the second intron of the hagfish vasotocin gene consists of 177 bp, the third intron has a length of ≈14.2 kbp, as deduced from the physical map (Fig. 1*A*). This intron is, thus, considerably larger than the introns found at the same relative positions in other members of the vertebrate vasopressin gene family (6, 29), which range in size from 90 to 1150 bp.

When the coding regions of the cDNA and genomic sequences are compared, a few divergences in the nucleotide and amino acid sequences are noted. These differences most likely represent allelic polymorphisms in the analyzed population. For instance, the cDNA differs from the genomic sequence by an A → G exchange at position 422 (cDNA numbering), which substitutes a glycine residue for an aspartate residue. In the 3′-untranslated region, the genomic sequence differs from that of the cDNA by the deletion of 3



B

```
VTcDNA  GGACAATAGTAAAGATACTGGCAGCGAAAGAGGAACGCTCTACAG----------------------------------------------------------------------------------------  45
VTGENE  ----------------------------------------------GATCCGACTCCTAGAATTTTTAAAAGTGATTCACCATTGAGAGATAGGACCAAAAACTATATATAGCACGACA  73

VTGENE  TCTCAAAATGTATTGGGCCCGATATGTTCTCGTTCACACTAATCGCTCAAAGTGATGTAATCATTTGGGTTTATTTAAGAATAATATAATTTTTGATTAAATTTTATAAGGTGGGAGGAG  193
VTGENE  GCGGTGGATAACTCAAAATGGTTCACACCCAACGCATCGCCATCCCCCTCCCTAGCGTCAGACGTCTTTGCCAAGTTTATCCACAGATCAAACACTGGGGATTTAATGGTGAAAAACCTC  313
VTGENE  TTAAAAGGAGCCATCTATTTACTGGGTACGTATAGTTCGCGGGCGGGGGGGGTGAAACAAGAAAAGGAAGGATGAAAAGAGAGAAAAAGGTGGGAAAAGGAATATGGAAGAAGTGTGGAG  433
VTGENE  ATAAGTCATTGGAATTTGGGAGGAGGTGGAAGTAGCTTGAGAGAAGGAAATGAAGTCTGGGACAAGATTGGAGTAAAGAGAAAGAAGGGAATGAAGTTTAGGAGAAAGGAGTAAAGAGAA  553
VTGENE  AGAAGGGAATGAAGTTTAGGAGAAAGGAGTAAAGAGAAAGAAGGGAATGAAGTTTCGGAGGAAATAGAAGCAAAGGGAGAGAAAAGGAAGGTGCATGTATGGAACGTCAGAGGGTGAGGG  673
VTGENE  GGTCAAAAAATTCGCTGCATTTGCGTACGTAAAAAATAGACGGTCTCATGAAGTCTTCGTCGTGTATTGTGGGAATTGGTATTGAACAATGATTAGTCAGGCCTCTACCTTTGCTGTTTA  793

                                                                                M  S  A  M  G  W  T  L  L  A  A  A  A  L  L  A
VTcDNA  ------------------------------------------------------------------GCTGCAATGAGTGCCATGGGATGGACGTTGCTAGCAGCAGCATTGCTTGCG  96
VTGENE  GCACAATGCTTTTAATTGCATGGTATATATATATATATATATATTATGTTTGAAAACTGTTGTGTATAGGCTGCAATGAGTGCCATGGGATGGACGTTGCTAGCAGCAGCATTGCTTGCG  913

        I  S  A  Q  S  N  G  C  Y  I  Q  N  C  P  R  G  G  K  R  A  V  E  T  E  L  H  S
VTcDNA  ATTAGCGCCCAGAGCAACGGCTGTTACATCCAGAACTGTCCCCGGGGTGGGAAACGAGCAGTGGAGACGGAGCTTCATTCT----------------------------------------  177
VTGENE  ATTAGCGCCCAGAGCAACGGCTGTTACATCCAGAACTGTCCCCGGGGTGGGAAACGAGCAGTGGAGACGGAGCTTCATTCTGTGAGTTCTGCATGCTTTCATTATCTGAGCTTGGCATGT  1033

VTGENE  TCCAACCCACTATGCAGTCATTTGCTTCAATGCACCTGATAATTCATTCTTTGTTCATGTAAAATCCAGCTATGTTCGATTTCAGTTCTGTACCCTCGTTAAAAAACACGATTTAAAAAAC  1153

                C  A  A  C  G  L  G  G  Q  C  V  G  P  S  I  C  C  G  G  L  L  G  G  G  R  G  G  G  C  I  V  G  G  P
VTcDNA  ------------------TGTGCAGCATGCGGACTGGGTGGCCAATGTGTGGGTCCATCCATCTGCTGCGGAGGATTGCTGGGAGGGGGCCGTGGAGGCGGTTGTATTGTTGGCGGACCG  279
VTGENE  ATGTATTTTCTCTACCAGTGTGCAGCATGCGGACTGGGTGGCCAATGTGTGGGTCCATCCATCTGCTGCGGAGGATTGCTGGGAGGGGGCCGTGGAGGCGGTTGTATTGTTGGCGGACCG  1273

        L  S  A  P  C  K  R  E  N  L  H  P  E  P  C  R  P  G  G  G  S  S  C  G  L  E  G  I  C  A  A  P  G  I  C  C  T  D  V
VTcDNA  TTGTCAGCACCCTGCAAACGTGAAAACCTCCATCCCGAGCCATGTAGACCTGGAGGGGGTTCATCTTGTGGCCTTGAGGGTATATGTGCAGCACCAGGCATCTGCTGCACTGACG-----  394
VTGENE  TTGTCAGCACCCTGCAAACGTGAAAACCTCCATCCCGAGCCATGTAGACCTGGAGGGGGTTCATCTTGTGGCCTTGAGGGTATATGTGCAGCACCAGGCATCTGCTGCACTGACGGTAAA  1393

VTGENE  TGGGGTAGGATATAGTT------- 6.2 kbp -------GGAGATTCAAAATAAATAAATACTCTACCGGTCAAAACTTTTAGAACATCCCCATTTTCCGAGTTTTTACTGAAATTTAA  7693
VTGENE  GCAGTTTAATGTTTCGGTATACTCTGAAATAAAAGCATAGAACAAACTAACAATTGCAGATTTAAAAAAAAATCATGGAATCGTTTTGTTGAACAAAATTTAATCGAAATTTTTGACTCA  7813
VTGENE  TCAAAGTAGCCACCTTTTGCAGATATAACAGCTGAACATACTCATGGCATTCTTTCTACAATGGAAATCAAATATTGTTCGGAAAGTTCTTCCCTACTCTGTTGCAGAAGTTCCCACAAA  7933
VTGENE  TGTGTTGCACTTGTAGGTTGCTTTGCTTTCACTCTTCTGTCCAGTTCATCCCAAACCAGCTCGATGGGGTTTAAGTCTGGAGACTGTGCTGGCCATTCCATGATTTGAAGCCTACCGTCT  8053
VTGENE  TGTTCTTCTCTTCTGAGGTCGTTCTGACATAGCCTGGAGGTATGTTTTGGGTCGTTATCTTGCTGTAGGATGAACCCCTGACCAACTAGGCGTAGACCAGAGGGTATTGCATGGCGCTGC  8173
VTGENE  AAAATGCTGTGGTAGCCCTTTCGGTTCAGGGTGCCAGTCACTCTGTGCAAGTCGCCGACTCTGGATCCAGCGAAAGAGCCCCAGATCATCAAGCTTCCTCCTCCATGTTTGACAGTTGGT  8293
VTGENE  GTCACACAATGAGGAACTATCCTTTTGCCTACTCAACAGCGTACAAAAACCCTGCGTGATGAAACGAAGATTTCAAATTTTGATTCATCGGTCCATAAGGCCTTCTTCCAGTTTTCAGTA  8413
VTGENE  GTCCACTGGCGGTGCTTCATGGCCCAGGCAAACCTCTTTTTCTTATTTTGCCATCTTCTAATTGTCTTCTGATTGAAACGAAGACACTCGTCTTCTGATTGTGTTGTAGCTTTGGGTCTG  8533
VTGENE  CCAGACCTCTTCCTGTCAGAGTTTCCTGCAGTTTCCAAGTGCCTTTTGATGGTGTAGGAAACTGTACTCGCTGACACCTTGGCTTCTTTGCAATTTCTCCAAAGGAAAGACCTACACTTT  8653
VTGENE  GAAGGGTTATAATGGTCTCTCTGTCTTCCTTTGTTAATTGTATTTTTCTCGCCATTATGATAGCAATATACTACTTCCTGCAGTGCAATATTGTCCAAATAATGTTTCAGAGGGTGCAGC  8773
VTGENE  AACACAATCTGTTCCAACACTGCTTTTTTACAGACAGAGGGTTTGTAACTAATCAACAAAAGTTGGGACACGTGGAGGAATTGTTTGCATCAACTTCCAAGGCTAAAATTACTTCCATCG  8893
VTGENE  TTGCAGAACAGCTGCAGTTGCGAACCCATTACTTGTTCTCTGAAAAGGGCCTTTTTTAACAACTCTGAAATTTATATTATATTTTTCAGTTTTTGGTAACCTTAACTTTTTTTTTTAACC  9013
VTGENE  TTTGCCAGTTTACCGCTTAGCTTTGTCCCATTTCAGGTTATTCACTGGACTTGAACTGCTTAAATTTCAGTAAAAACTGGAAAAATGGGGGTGTTCTAAAACTTTCGACCGGTAGAGTAA  9133

                                                                                                                T  C  S  I  D  A  T  C  D
VTcDNA  ----------------------------------------------------------------------------------------------------TCACTTGCTCCATCGATGCCACCTGCGAT  423
VTGENE  ATAAGGTTAGGCTATAAAA---------- 6.45 kbp ----------CGATTGTCTTAGATATAAAATCTTTGCTTCCTTGAACAGTCACTTGCTCCATCGATGCCACCTGCGGT  15693
                                                                                                                                G
        D  V  T  E  K  A  G  V  T  F  S  G  A  T  G  G  L  A  N  P  T  G  D  L  L  R  K  V  L  L  L  A  N  A  D  L  E  *
VTcDNA  GATGTCACAGAGAAAGCTGGTGTGACCTTCTCCGGTGCCACTGGTGGATTGGCTAACCCAACTGGCGACTTGCTTCGGAAAGTTTTACTCCTGGCCAATGCTGACTTGGAATAATAATAA  543
VTGENE  GATGTCACAGAGAAAGCTGGTGTGACCTTCTCCGGTGCCACTGGTGGATTGGCTAACCCAACTGGCGACTTGCTTCGGAAAGTTTTACTCCTGGCCAATGCTGACTTGGAATAATAATAA  15813

VTcDNA  TAATAATAATAATAATAATAATGATTGTGTTTAGCTGTGTCACTTTACCTGTGTTCTGCCACCTATTTGTATCACGTCACTTCTTGTGTGCCTTTCAGGCGCTCTTGGTTATATTTTCCC  663
VTGENE  TAATAATAATAAT---------GATTGTGTTTAGCTGTGTCACTTTACCTGTGTTCTGCCACCTATTTGTATCACGTCACTTCTTGTGTGCCTTTCAGGCGCTCTTGGTTATATTTTCCC  15924

VTcDNA  CGAAGTACAAGCTACAATTTAAAACTTATTCAA-CCCTTGCAGTCACAGTGTAA--GCCTTAATTCCTAACACTGACTCCAACTACAAACTCTCTTCAAAACAACCACGACAGCTTTTTG  780
VTGENE  CGAAGTACAAGCTACAATTTAAAACTTATTCAAACCCTTGCAGTCACAGTGTAAAAGCCTTAATTCCTAACACTGACTCCAACTACAAACTCTCTTCAAAACAACCACGACAGCTTTTTG  16044

VTcDNA  CATTGGTTGGACTAGTTTTAAGTTTGCAGGAATTTTTCAGACATTTCTTTCAATGAAGGCCAAGAATGTTGAATGTCTCATATGAAGTCCATTGGCAGACGTGCACGGACGTTTTACGGT  900
VTGENE  CATTGGTTGGACTAGTTTTAAGTTTGCAGGAATTTTTCAGACATTTCTTTCAATGAAGGCCAAGAATGTTGAATGTCTCATATGAAGTCCATTGGCAGACGTGCACGGACGTTTTACGGT  16164

VTcDNA  GACGTCACATTCCACTACCCACGACATGATTTCTTGCCTTTTTATATTGTGAAAGTTGTTTTTGATTGACTGTACAACCTTGTCTTGCTCTTGTGAAAAATCATTGCAAACAATGACCGT  1020
VTGENE  GACGTCACATTCCACTACCCACGACATGATTTCTTGCCTTTTTATATTGTGAAAGTTGTTTTTGATTGACTGTACAACCTTGTCTTGCTCTTGTGAAAAATCATTGCAAACAATGACCGT  16284

VTcDNA  TAGAAAATAAATGAAGGATCATGAATGCG  1049
VTGENE  TAGAAAATAAATGAAGGATC  16304
```

FIG. 1. (*A*) Structure and sequence analysis of the hagfish vasotocin cDNA and gene. Scale units of the diagram are kbp for the gene and 100 bp for the cDNA. (*Top*) Alignment of three independent vasotocin genomic clones. Orientation of inserts is indicated by positions of the T3 and T7 RNA polymerase promoters contained in the λDASHII cloning vector. (*Middle*) Physical map of the vasotocin gene. Filled boxes indicate vasotocin exons. Stippled box represents location of Tes*1* element, the orientation of whose putative transposase reading frame is indicated by an arrow. The star indicates an *Eco*RI site, present only in clones λVTgene1 and λVTgene2. (*Bottom*) cDNA structure, where boldface lines represent nontranslated regions. Boxes denote different moieties of the vasotocin precursor. Boldface letters designate exons that code for different parts of the precursor. S, signal peptide; V, vasotocin; NP, neurophysin. (*B*) Comparison of cDNA and genomic DNA sequences. Numbers at right indicate nucleotide positions. Exonic sequences are shown in plain text; intronic sequences are in italic. The deduced amino acid sequence of the vasotocin (VT) precursor is shown. The genomic sequence predicts the same polypeptide sequence, except for a glycine instead of an aspartate; this difference results from a polymorphism at nucleotide 15692. The Tes*1* transposon sequence is shown in the middle part, where underlining indicates the inverted terminal repeats. Note that the putative transposase is encoded by the complementary strand. The approximate sizes of those genomic sequences, yet to be determined, are indicated. The polyadenylylation signal at the end of the cDNA and the gene is also underlined.

out of 11 triplets in the stop-codon cluster and the insertion of a total of 3 adenine residues (Fig. 1).

Identification of the Hagfish Tes*1* Transposon. The large third vasotocin intron could easily accommodate an additional gene(s). To identify possible transcribed regions within this intron, the three differently sized genomic vasotocin clones were digested with restriction enzymes, Southern blotted, and hybridized with labeled hagfish brain poly(A)$^+$

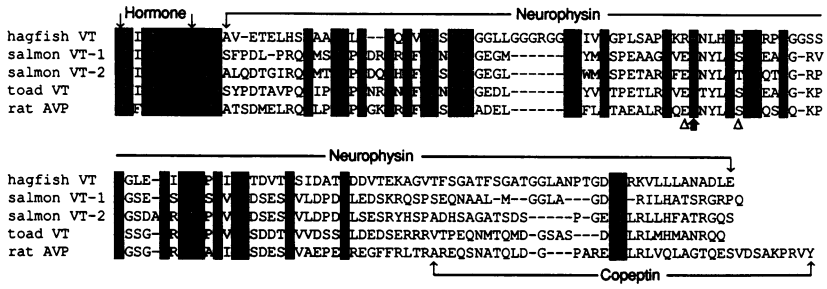


FIG. 2. Alignment of vasopressin family prohormone sequences from different vertebrate classes. Only one representative of each class is shown, except for teleosts that contain two highly divergent precursors. Dark boxes indicate residues that are conserved among all compared sequences. The solid arrow points to a conserved glutamic acid residue, essential for the noncovalent binding of vasopressin-like peptides by the neurophysin (25, 26). Open arrowheads indicate residues that form antiparallel hydrogen bonds in bovine neurophysin II dimers (26). Sequences are taken from refs. 6, 7, and 9. VT, vasotocin; AVP, arginine vasopressin.

RNA. Several hybridizing bands were obtained, but not all fragments identified by restriction mapping hybridized to labeled hagfish brain cDNA (Fig. 3*B*, closed arrowheads). At the stringent conditions chosen, no hybridization signals were seen when using the vasotocin cDNA clone (Fig. 3*B*, lane VT cDNA) as a control, which is due to low levels of vasotocin mRNA in total brain poly(A)$^+$ RNA preparations and which agrees with the results of RNA blots that yield a hybridization signal with mesencephalon but not with total brain RNA (Fig. 3*A*).

The shortest hybridizing restriction fragment—a 1.1-kbp *Hin*dIII fragment—was subsequently sequenced and used for a data base search (National Biomedical Research Foundation-Protein Identification Resource data base, release 28). The longest ORF of this fragment is located on the opposite DNA strand to that encoding the vasotocin precursor. This ORF predicts a protein sequence strikingly similar to putative transposases of Tc*1*-like elements from nematodes and *Drosophila* (10–14). Further sequencing up and downstream of the *Hin*dIII fragment revealed a stretch of 66 bp of imperfect inverted terminal repeats flanking the region encoding the putative transposase (Fig. 4). Structural features, such as inverted terminal repeats and a putative transposase, are typical of nonretroviral mobile genetic elements (30), indicating that the third vasotocin intron sequence contains a transposon. When this transposon of ≈1.5 kbp in size, termed Tes*1*, was labeled and used to hybridize with the original Southern blot, no bands other than those marked by open arrowheads were detected (Fig. 3*B*), suggesting that

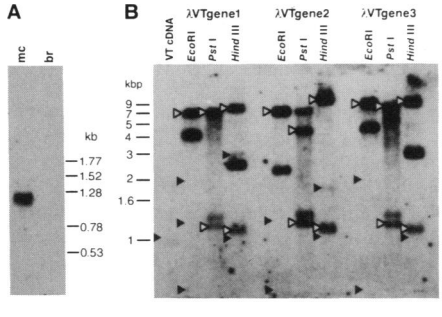


FIG. 3. (*A*) RNA blot analysis of the hagfish vasotocin mRNA. Ten micrograms of poly(A)$^+$ RNA from dissected hagfish mesencephali (mc) or total brains (br) was separated in a 1.3% (wt/vol) agarose gel and hybridized with a labeled *Sma* I/*Eco*RI restriction fragment of the vasotocin cDNA (nucleotides 140–407). Positions of marker bands (low-molecular-mass RNA ladder; BRL) are indicated. (*B*) Identification of Tes*1* transposon within the hagfish vasotocin gene. One microgram of DNA of genomic clones, digested as indicated, or the vasotocin cDNA as a control, was loaded per lane, separated in a 1% (wt/vol) agarose gel, and transferred onto a Hybond-N membrane, which was subsequently hybridized for 36 hr with 1 μg of labeled cDNA from total brains. After high-stringency washes the membrane was exposed to Kodak XAR-5 film at −80°C for 3 days. Positions of size markers (1-kbp ladder, BRL) are indicated. Open arrowheads indicate hybridizing restriction fragments that contain Tes*1* transposon sequences. Filled arrowheads denote nonhybridizing restriction fragments of the cloned sequences.

additional Tes*1* copies are absent from the third intron. The additional hybridizing bands unrelated to Tes*1* shown in Fig. 3*B* have not yet been characterized.

Similarity of the hagfish Tes*1* putative transposase to its counterparts from nematodes and flies is restricted to ≈200 amino acid residues (Fig. 4); this region exhibits ≈40% identity to the nematode transposases (*Caenorhabditis elegans* Tc*1* element, ref. 10; *Caenorhabditis briggsae* Tcb*1* element, ref. 11), and between 28 and 32% identity to *Drosophila* transposases (Uhu, ref. 13; Minos, ref. 14; and HB-1, refs. 11 and 12). The ATG codon assigned as the initiator codon (Fig. 4) predicts a considerably larger transposase compared with that of Tc*1* (407 vs. 273 amino acid residues); however, a definite prediction of the structure of the Tes*1* transposase must await the sequence analysis of further Tes*1* copies because the presented transposase reading frame contains three in-frame stop codons, a feature shared with many elements of other transposase families (31). Transposition of such nonautonomous elements depends on the presence of trans-acting functional transposases generated from complete copies (31).

Southern and RNA blot analysis of hagfish genomic DNA and brain poly(A)$^+$ RNA, respectively, clearly demonstrated the presence of multiple copies of the Tes*1* transposon within the hagfish genome and of Tes*1*-like elements in numerous other hagfish genes expressed in brain (data not shown). Whether any of these are autonomous elements—i.e., able to encode functional transposases, remains to be determined.

Eukaryotic mobile genetic elements have been classified according to their mode of transposition (30). Class I elements, often found in vertebrates, resemble retroviral prosequences and move within their host genomes via RNA intermediates. In contrast, class II elements, characterized by short inverted terminal repeats flanking putative transposase-encoding reading frames, transpose directly from DNA to DNA. Although several distinct families of class II transposable elements are known in invertebrates and plants (30, 31), only a few candidate inverted terminal-repeat transposons have to date been identified in vertebrate genomes, where class I elements seem to predominate. However, all vertebrate class II elements described (32) lack transposase-encoding reading frames. Because the cloned hagfish Tes*1* element contains a reading frame that encodes a protein reminiscent of a transposase, this demonstrates the presence of a typical inverted terminal-repeat transposon within vertebrates. The presence of Tc*1*-like elements in nematodes (10, 11), flies (11–14), and cyclostomes indicates that they represent members of the most widely dispersed class II transposon family currently known in the animal kingdom. It remains to be established whether this broad distribution is due to evolutionary conservation or to the horizontal transfer of these elements across species borders, which has been proposed for other transposon families (33, 34).

Tc*1* elements are the major cause of spontaneous mutations in certain *C. elegans* strains (35) and have successfully been used as probes for the cloning of several genes (35). Unfortunately, the hagfish is probably one of the least studied

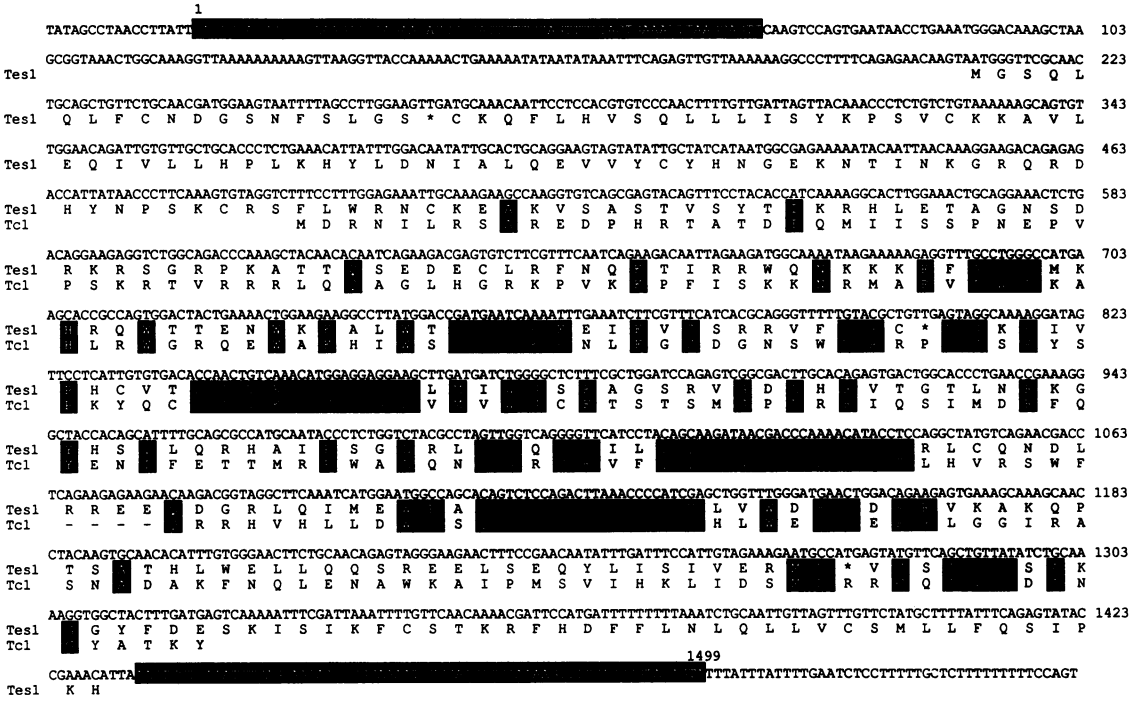


FIG. 4. Sequence of the hagfish Tes*1* element and flanking nucleotides. Within the DNA sequence, boxes denote the inverted terminal repeats. The predicted amino acid sequence of an ORF, which is interrupted by three in-frame stop codons (indicated by stars), is shown below the DNA sequence. Identical amino acid residues in the putative Tes*1* and Tc*1* (ref. 10) transposases are indicated by dark boxes. Dashes denote gaps that were introduced to maximize alignment.

organisms in genetic terms; this is emphasized by the fact that only a single sequence, other than those described here, is currently known. It is, thus, premature to comment on the mutational impact of Tes*1* elements in the hagfish genome.

We thank David Ko (Calgary) for collecting adult Pacific hagfish specimens at the Banfield Marine Station, Vancouver Island, BC, Canada; Dr. Yuji Okawara (Calgary) for preparing the hagfish mesencephalon RNA; Werner Rust and Sönke Harder for technical assistance; Dr. Steven D. Morley for advice on cDNA library construction; and Drs. Mark Darlison and Wolfgang Meyerhof for critically reading the manuscript. This project was supported by grants from the Deutsche Forschungsgemeinschaft to D.R. and the Medical Research Council of Canada to K.L. The data presented are part of a thesis by J.H.

1. Carroll, R. L. (1988) in *Vertebrate Paleontology and Evolution* (Freeman, New York), pp. 39–41.
2. Nieuwenhuys, R. (1972) *J. Comp. Neurol.* **145,** 165–178.
3. Grillner, S., Wallén, P. & Brodin, L. (1991) *Annu. Rev. Neurosci.* **14,** 169–199.
4. Acher, R. (1980) *Proc. R. Soc. London* B **210,** 21–43.
5. Maetz, J. & Lahlou, B. (1974) in *Handbook of Physiology*, eds. Knobil, E. & Sawyer, W. H. (Am. Physiol. Soc., Washington), Vol. 4, pp. 521–544.
6. Richter, D. (1987) in *The Peptides*, eds. Udenfriend, S. & Meienhofer, J. (Academic, New York), Vol. 8, pp. 41–75.
7. Nojiri, H., Ishida, I., Miyashita, E., Sato, M., Urano, A. & Deguchi, T. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 3043–3046.
8. Heierhorst, J., Morley, S. D., Figueroa, J., Krentler, C., Lederis, K. & Richter, D. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 5242–5246.
9. Heierhorst, J., Mahlmann, S., Morley, S. D., Coe, I. R., Sherwood, N. M. & Richter, D. (1990) *FEBS Lett.* **260,** 301–304.
10. Rosenzweig, B., Liao, L. W. & Hirsh, D. (1983) *Nucleic Acids Res.* **11,** 4201–4209.
11. Harris, L. J., Baillie, D. L. & Rose, A. M. (1988) *Nucleic Acids Res.* **16,** 5991–5998.
12. Henikoff, S. & Plasterk, R. H. A. (1988) *Nucleic Acids Res.* **16,** 6234.
13. Brezinsky, L., Wang, G. L., Humphreys, T. & Hunt, J. (1990) *Nucleic Acids Res.* **18,** 2053–2059.
14. Franz, G. & Savakis, C. (1991) *Nucleic Acids Res.* **19,** 6646.
15. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18,** 5294–5299.
16. Aviv, H. & Leder, P. (1972) *Proc. Natl. Acad. Sci. USA* **69,** 1408–1412.
17. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning* (Cold Spring Harbor Lab., Cold Spring Harbor, NY), 2nd ed.
18. Gubler, U. & Hoffman, B. J. (1983) *Gene* **25,** 263–269.
19. Short, J. M., Fernandez, J. M., Sorge, J. A. & Huse, W. D. (1988) *Nucleic Acids Res.* **16,** 7563–7600.
20. Wood, W. I., Gitschier, J., Lasky, L. & Lawn, R. M. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1585–1588.
21. Feinberg, A. P. & Vogelstein, B. (1983) *Anal. Biochem.* **132,** 6–13.
22. Yanish-Perron, C., Vieira, J. & Messing, J. (1985) *Gene* **33,** 103–119.
23. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
24. Wahl, G. R. (1989) *Stratagies* **2,** 1–3.
25. Burman, S., Wellner, D., Chait, B., Chaudhary, T. & Breslow, E. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 429–433.
26. Chen, L., Rose, J. P., Breslow, E., Yang, D., Wen-Rui, C., Furey, W. F., Sax, M. & Wang, B.-C. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 4240–4244.
27. Fassina, G. & Chaiken, I. M. (1988) *J. Biol. Chem.* **263,** 13539–13543.
28. Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472.
29. Morley, S. D., Schönrock, C., Heierhorst, J., Figueroa, J., Lederis, K. & Richter, D. (1990) *Biochemistry* **29,** 2506–2511.
30. Finnegan, D. J. (1990) *Curr. Opin. Cell Biol.* **2,** 471–477.
31. Gierl, A. & Frey, M. (1991) *Curr. Opin. Genet. Dev.* **1,** 494–497.
32. Carroll, D. C., Knutzon, D. S. & Garrett, J. E. (1989) in *Mobile DNA*, eds. Howe, M. & Berg, D. (Am. Soc. Microbiol. Publ., Washington), pp. 567–574.
33. Calvi, B. R., Hong, T. J., Findley, S. D. & Gelbart, W. M. (1991) *Cell* **66,** 465–471.
34. Simonelig, M. & Anxolabéhère, D. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 6102–6106.
35. Moerman, D. G. & Waterston, R. H. (1989) in *Mobile DNA*, eds. Howe, M. & Berg, D. (Am. Soc. Microbiol. Publ., Washington), pp. 537–556.