Origin of Gene Overlap: The Case of TCP1 and ACAT2

Seikou Shintani, Colm O’huigin, Satoru Toyosawa, Věra Michalová and Jan Klein

Max-Planck-Institut für Biologie, D-72076 Tübingen, Germany

Manuscript received December 10, 1998
Accepted for publication February 11, 1999

ABSTRACT

The human acetyl-CoA acetyltransferase 2 gene, ACAT2, codes for a thiolase, an enzyme involved in lipid metabolism. The human T-complex protein 1 gene, TCP1, encodes a molecular chaperone of the chaperonin family. The two genes overlap by their 3'-untranslated regions, their coding sequences being located on opposite DNA strands in a tail-to-tail orientation. To find out how the overlap might have arisen in evolution, the homologous genes of the zebrafish, the African toad, caiman, platypus, opossum, and wallaby were identified. In each species, standard or long polymerase chain reactions were used to determine whether the ACAT2 and TCP1 homologs are closely linked and, if so, whether they overlap. The results reveal that the overlap apparently arose during the transition from therapsid reptiles to mammals and has been retained for >200 million years. Part of the overlapping untranslated region shows remarkable sequence conservation. The overlap presumably arose during the chromosomal rearrangement that brought the two unrelated and previously separated genes together. One or both of the transposed genes found by chance signals that are necessary for the processing of their transcripts to be present on the noncoding strand of the partner gene.

THE classical notion of genes being arranged on chromosomes in a bead-on-string-like fashion had to be amended when the first genomic sequences, those of viruses, became known. One of the first such sequences, that of the bacteriophage φX174 (Sanger et al. 1977), revealed that in addition to the tandemly arranged genes, there were two pairs of genes in which different proteins were translated in two reading frames from a common DNA sequence (Barrell et al. 1976). This observation was the first documented instance of overlapping genes. Many other instances of overlapping genes and other deviations from strictly tandem gene arrangements in the genomes have been described since then (Normark et al. 1983). In some cases, a gene of a given pair is nested within another gene, as in the two φX174 pairs; in others, the two genes overlap partially.

Overlapping genes occur frequently in viral genomes as well as in genomes of cellular prokaryotes and prokaryote-derived organelles such as mitochondria (Normark et al. 1983). They occur less frequently in nuclear genomes of eukaryotes (Williams and Fried 1986; Adelman et al. 1987; Emi et al. 1988; Morel et al. 1989; Cawthon et al. 1991; Lauden et al. 1991; Grima et al. 1992; Shaiqy and Avadhani 1992; Ashworth 1993; Nicoloso et al. 1994; Aaronson et al. 1996; Hadano et al. 1996; Swalla and Jeffery 1996; Van Bokhoven et al. 1996; Cooper et al. 1998; Joseph 1998). There are two principal types of overlapping genes: in one type, the transcribed (and translated) reading frames of the genes are on the same DNA strand; in the other, they are on complementary strands. The former type is less common than the latter, at least in nuclear genomes.

Little is known about the manner in which the overlap arises during evolution. Though overlaps have been identified, less effort has been expended on determining their origins. We consider two ways by which, theoretically, overlapping genes can come into existence. Within a gene-constituting DNA stretch, often more than one reading frame, either on the same or on complementary strands, can potentially code for a peptide. If an initiation codon and a transcription initiation site arise by chance within the stretch and in register with the extra open reading frame, two or more mRNA types may be transcribed from the same locus. Alternatively, two independently derived genes on the same or on different chromosomes can be brought together, for example, by translocation, and arranged in such a way that each derives part of its transcript from the same or complementary DNA sequence as the other.

To investigate the mode of origin of overlapping eukaryotic genes, we chose the ACAT2-TCP1 pair. The ACAT2 or acetyl-CoA acetyltransferase 2 gene codes for an enzyme of lipid metabolism, a member of the thiolase family (Clinkenbeard et al. 1973; Middleton 1973, 1974; Song et al. 1994). Thiolases, a widely distributed group of enzymes found in both prokaryotes and eukaryotes, are of two basic types: 3-ketoacyl-CoA thiolases (type I, EC 2.3.1.16) and acetoacetyl-CoA thiolases (type II, EC 2.3.1.9, see Middleton 1975; Igual et al. 1992). The former are involved in the degradation of fatty acids by β-oxidation; the latter participate in the biogenesis of

Corresponding author: Jan Klein, Max-Planck-Institut für Biologie, Abteilung Immunregnetik, Corrensstr. 42, D-72076 Tübingen, Germany. E-mail: jan.klein@tuebingen.mpg.de

Genetics 152: 743–754 (June 1999)
steroids and the formation of ketone bodies. Type I thiolases have a broad chain-length specificity; type II enzymes are specifically involved in the thiolysis of acetyl-CoA. Mammalian genomes contain at least five thiolase-encoding loci that specify three mitochondrial type I thiolases (Fukao et al. 1990; Abe et al. 1993; Kamijo et al. 1994), one peroxisomal type I thiolase (Fairbairn and Tanner 1989), and one cytosolic type II thiolase (Song et al. 1994). In humans, the locus encoding the cytosolic thiolase is designated ACAT2; the encoded enzyme catalyzes the condensation of two acetyl-CoA molecules into acetoacetyl-CoA, which is then converted via several steps into ketones (Middleton et al. 1974). The rat ACAT2 product is a homotetramer that is abundantly expressed in the liver, brain, and adrenals and poorly expressed in most other tissues (Middleton et al. 1974).

The TCP1 or T-complex protein 1 gene codes for a molecular chaperone that assists in the folding of proteins during their synthesis or their recovery from a denatured state (Willison et al. 1986; Ellis and van der Vies 1991; Horwich and Willison 1993). It is also a member of a large family of proteins, specifically a class of molecular chaperones known as chaperonins. (Yaffe et al. 1992; Horwich and Willison 1993; Kubota et al. 1994). This class includes GroEL of Escherichia coli, the mitochondrial heat shock protein Hsp60, the plastid Rubisco subunit-binding protein, and the archaean protein TF55. Chaperonins are involved in the folding, transport, and assembly of newly synthesized proteins that, in the case of TCP1-containing chaperonins, include actin and tubulin. In mammals, the TCP1 gene codes for one subunit ($\alpha$) of a particle that contains at least six other subunit types ($\beta$, $\gamma$, $\delta$, $\epsilon$, $\zeta$, and $\eta$) that are all encoded in distinct but related genes (Kubota et al. 1994). Because the TCP1 gene has apparently nothing to do with the T-complex phenotype, it has been renamed CCTA, for chaperonin-containing TCP1$\alpha$ (Kubota et al. 1994). In the mouse, the TCP1 gene is expressed in several tissues, but most abundantly in the testes (Willison et al. 1986).

In mice and humans, the ACAT2 and TCP1 genes are located in the same chromosomal regions on chromosomes 17 and 6q25.3-q26, respectively (Willison et al. 1987; Ashworth 1993; Masuno et al. 1996). In both species, the two genes overlap in a manner shown in Figure 1. The coding sequences of the two genes are located on opposite DNA strands and are in a tail-to-tail orientation to each other. They share a DNA segment encompassing portions of their 3′-untranslated regions (UTRs) and, in one direction, also part of the translated region. In the mouse, the pairs have apparently undergone a tandem duplication so that the genes are arranged in the order TCP1 ... ACAT2 ... TCP1 ... ACAT2. (One of the mouse ACAT2 genes was originally erroneously designated TCP1x, see Dudley et al. 1991.) The aim of the present study was to determine at which stage in vertebrate evolution and in what manner the overlap of the ACAT2 and TCP1 genes arose.

### MATERIALS AND METHODS

#### Source and isolation of DNA:

The spleen of an adult red-necked wallaby (Macropus rufogriseus) was obtained from an animal that died in the Hamburg-Hagenbeck Zoological Garden. Tissues from the gray short-tailed opossum (Monodelphis domestica) were obtained from the colony maintained by Professor W. H. Stone (Department of Biology, Trinity University, San Antonio, Texas). DNA from the duck-billed platypus (Ornithorhynchus anatinus) was provided by Dr. Robert W. Slade (Queensland Institute for Medical Research, Royal Brisbane Hospital, Australia). Fertilized eggs of a smooth-fronted caiman (Paleosuchus palpebrosus) were obtained from Dr. Hans-Peter Herrmann (Köln Zoo, Germany). African clawed toads (Xenopus laevis) were provided by Dr. C. Dreyer (the Max Planck Institute for Developmental Biology, Tübingen, Germany). Zebras (Danio rerio) bred in our aquarium were used. All tissue samples were kept frozen at −70°C until their use. Genomic DNA was isolated from the tissues by phenolchloroform extraction.

#### cDNA library construction and screening:

Animals were killed under anesthesia, and their tissues were removed and frozen in liquid nitrogen. The frozen tissues were homogenized to a fine powder, and total RNA was extracted. Poly(A)$^+$ RNA isolation and cDNA synthesis were performed with the help of the mRNA purification kit (Pharmacia Biotech, Freiburg, Germany) and the TimeSaver cDNA synthesis kit (Pharmacia Biotech), respectively. The cDNA was inserted into the EcoRI-digested a gt10 vector (Stratagene, Heidelberg, Germany), and the cDNA library was in vitro packaged with the help of the Gigapack cloning kit (Stratagene) and used to transform competent E. coli M 1544 bacteria. The opossum, caiman, toad, and zebrafish libraries were amplified once to titers of $4.0 \times 10^9$, $1.5 \times 10^9$, $1.8 \times 10^9$, and $1.0 \times 10^{11}$ pfu, respectively.

#### PAC clone screening:

PAC zebrafish library filters (library BUSMP706) were obtained from the Resource Center of the German Human Genome Project at the Max-Planck-Institut for Molecular Genetics (Berlin). Probes for TCP1 were prepared by polymerase chain reaction (PCR) amplification of the zebrafish library using the zebrafish-specific primers Tcp F5 and Tcp F7, as well as Acat F12 and Acat F14, which yielded products of ~600 and 450 bp, respectively, covering the bulk of the translated regions of the corresponding gene transcripts. Filters were hybridized at 65°C in 7% sodium dodecylsulfate (SDS), 0.5 m sodium phosphate, pH 7.2, and 1 mm ethylenediaminetetraacetic acid (EDTA), and were washed twice in 40 mm sodium phosphate containing 0.1% SDS. Positive hybridization signals from two ACAT2-containing PAC clones (numbers G1276Q2 and G23224), as well as two TCP1-containing clones (numbers H0274Q2 and O23263), were confirmed by PCR amplification of the probe regions from the PAC clones.

#### PCR amplification:

ACAT2 sequences of the opossum, caiman, toad, and zebrafish were amplified from cDNA libraries using primers based on a comparison of human and mouse sequences (Table 1) in combination with a vector primer in an anchored PCR. Genomic DNA or lysate of the cDNA libraries (1 µl) was amplified by PCR in 50 µl PCR buffer (1.5 mm MgCl₂, 200 µm dNTP, 10 mm Tris buffer, pH 8.5) in the presence of the two primers and 2.5 units of Taq polymerase (Pharmacia Biotech). Amplification was performed in the PT C-100 Thermal Cycler (MJ Research Inc., Watertown, MA). After the first cycle at 95°C for 3 min, 35 cycles followed, each consisting of 1 min denaturation at 95°C, 1 min annealing at...
**Overlapping Genes**

**Human**

```
<table>
<thead>
<tr>
<th>ACAT2</th>
<th>TCP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>195 bp</td>
<td>215 bp</td>
</tr>
<tr>
<td>67 bp</td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>PAS</td>
</tr>
<tr>
<td>21 bp</td>
<td>16 bp</td>
</tr>
</tbody>
</table>
```

**Mouse**

```
<table>
<thead>
<tr>
<th>ACAT2</th>
<th>TCP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>183 bp</td>
<td>736 bp</td>
</tr>
<tr>
<td>113 bp</td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>PAS</td>
</tr>
<tr>
<td>21 bp</td>
<td>16 bp</td>
</tr>
</tbody>
</table>
```

**Opossum**

```
<table>
<thead>
<tr>
<th>ACAT2</th>
<th>TCP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>254 bp</td>
<td>266 bp</td>
</tr>
<tr>
<td>133 bp</td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>PAS</td>
</tr>
<tr>
<td>16 bp</td>
<td></td>
</tr>
</tbody>
</table>
```

**Wallaby**

```
<table>
<thead>
<tr>
<th>ACAT2</th>
<th>TCP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>246 bp</td>
<td>263 bp</td>
</tr>
<tr>
<td>133 bp</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PAS</td>
</tr>
<tr>
<td></td>
<td>16 bp</td>
</tr>
</tbody>
</table>
```

**Platypus**

```
<table>
<thead>
<tr>
<th>ACAT2</th>
<th>TCP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>237 bp</td>
<td>253 bp</td>
</tr>
<tr>
<td>87 bp</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PAS</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Figure 1.—Diagram depicting the ACAT2-TCP1 overlapping region in different vertebrate species. The human and mouse diagrams are based on the data of Ashworth (1993), Song et al. (1994), Dudley et al. (1991), and Kubota et al. (1994). The diagrams for the remaining species are deduced from the data described in the present study. For each species, the two strands of genomic DNA are depicted, the last exon of the coding strand is shown as a rectangle, the 3' untranslated region as a thick line, and the rest of the strand as a thin line. The distances between the stop codons of the two genes (dotted line), between the stop codon and the conserved polyadenylation signal (PAS), as well as between the PAS and the cleavage site (CS) of the transcript are indicated in base pairs. Alternative PAS, which are known, for example, in humans (Willison et al. 1997), are not indicated.

The annealing temperature, and 2 min extension at 72°C. The final extension was for 10 min at 72°C. Long PCR was carried out with the help of the GeneAmp XL PCR Kit (Perkin Elmer Applied Biosystems, Freiburg, Germany) in the GeneAmp PCR system 9600 (Perkin Elmer-Cetus, Norwalk, CT) and consisted of one cycle at 94°C for 30 sec, followed by 12 cycles, each for 30 sec at 94°C and 10 min at 64°C. In the next 24 cycles, the reaction time at 64°C was extended by 15 sec in every cycle; the reaction was completed by a final primer extension for 10 min at 72°C.

**Cloning and sequencing:** Twenty microliters of the PCR amplification product was purified by electrophoresis in 1.5% low-melting-point agarose (GIBCO BRL, Eggenstein, Germany) and the band was identified by ethidium bromide staining, excised, and isolated from the gel using the QIAEX extraction kit (Hilden, Germany). The isolated DNA was blunt ended, phosphorylated, and ligated to Smal-digested pUC18 plasmid vector with the SureClone ligation kit (Pharmacia Biotech). The reaction products were transformed into competent E. coli XL-1 blue bacteria by standard methods and plated on LB agar containing ampicillin (50 μg/ml). Transformsants were grown overnight in LB broth containing ampicillin, and minipreps were prepared according to the standard Qiagen protocol. Two to five micrograms of DNA was used in the dideoxy sequencing reactions with the AutoRead Sequencing kit (Pharmacia Biotech). The reactions were processed by the automated laser fluorescent sequencer (Pharmacia Biotech). The GenBank accession nos. of the sequences are AF143488-AF143500.

**Data analysis:** The nucleotide sequences were aligned with the aid of the SeqPup computer program (Gilbert 1995). Sequence similarities were evaluated with the aid of the DottyPlot computer program (Gilbert 1989). Substitution rates were estimated by the method of Li (1993) from sequences
TABLE 1

<table>
<thead>
<tr>
<th>Designation</th>
<th>Sequence</th>
<th>Orientation</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acat 4</td>
<td>5'-ACTGGTACAATCTTTGTCAAA-3'</td>
<td>A</td>
<td>E3</td>
</tr>
<tr>
<td>Acat 7</td>
<td>5'CCATCATTTATTCTGAAGCATT-3'</td>
<td>A</td>
<td>E4/E3</td>
</tr>
<tr>
<td>Acat C1</td>
<td>5'TATTGCCCTTGAGCACCCCTCT-3'</td>
<td>S</td>
<td>E7</td>
</tr>
<tr>
<td>Acat F4</td>
<td>5'TTGGACAGGGAATGGTGCAGT-3'</td>
<td>S</td>
<td>E3</td>
</tr>
<tr>
<td>Acat F8</td>
<td>5'GGCATCTCTCTTGTGCTATCTTCCT-3'</td>
<td>S</td>
<td>E7</td>
</tr>
<tr>
<td>Acat F12</td>
<td>5'CACACATAGCGATGCCATTCTC-3'</td>
<td>A</td>
<td>E7</td>
</tr>
<tr>
<td>Acat F14</td>
<td>5'GGTACTGTGACAGCTGGTAATG-3'</td>
<td>S</td>
<td>E5</td>
</tr>
<tr>
<td>Acat G1</td>
<td>5'AGGCATTGCGCYGKTGCATC-3'</td>
<td>S</td>
<td>E7</td>
</tr>
<tr>
<td>Acat G3</td>
<td>5'TTGGACAGCCTTGGACACCCTCT-3'</td>
<td>S</td>
<td>E7</td>
</tr>
<tr>
<td>Acat O4</td>
<td>5'CTGGCATTCTTTGTCAAA-3'</td>
<td>S</td>
<td>E3</td>
</tr>
<tr>
<td>Acat X1</td>
<td>5'CCCTCTGAGCTTCTGGTGC-3'</td>
<td>S</td>
<td>E3</td>
</tr>
<tr>
<td>Acat Xa1</td>
<td>5'CCCTCTGGAATGCTCTGTGTTGC-3'</td>
<td>S</td>
<td>E7</td>
</tr>
<tr>
<td>Tcp C1</td>
<td>5'ATTGGGCTTGACTTAATAAATGGAA-3'</td>
<td>S</td>
<td>E12</td>
</tr>
<tr>
<td>Tcp F4</td>
<td>5'ATTACCATCTTCCGAATGATGCTCA-3'</td>
<td>S</td>
<td>E12</td>
</tr>
<tr>
<td>Tcp F5</td>
<td>5'GATTGGACGGCGTCCTGATA-3'</td>
<td>A</td>
<td>E12</td>
</tr>
<tr>
<td>Tcp F7</td>
<td>5'GGTTGTGCTTCTTGCTGGTCA-3'</td>
<td>S</td>
<td>E9</td>
</tr>
<tr>
<td>Tcp G1</td>
<td>5'TGTGCATYGGCGGAGGGAT-3'</td>
<td>S</td>
<td>E12</td>
</tr>
<tr>
<td>Tcp G3</td>
<td>5'AAACAGAGGCTTTTGATGGAA-3'</td>
<td>S</td>
<td>E12</td>
</tr>
<tr>
<td>Tcp G8</td>
<td>5'AGGCCTGCTATTGCAGAGTTCG-3'</td>
<td>S</td>
<td>E11</td>
</tr>
<tr>
<td>Tcp Xa1</td>
<td>5'TAAACAAGCAGGCTTGGGTCTTC-3'</td>
<td>S</td>
<td>E12</td>
</tr>
</tbody>
</table>

A, antisense; S, sense; E, exon.

RESULTS

Human ACAT2 gene organization: As a prelude to the main study, we determined the exon-intron borders of the human ACAT2 gene to better plan experiments concerned with the overlap of the ACAT2 and TCP1 genes and interpret their results. To this end, we used the published cDNA sequences (Song et al. 1994) to synthesize a series of primer pairs corresponding to stretches spaced at short intervals along the sequence. In each pair, the sense and antisense primers corresponded to cDNA sequence stretches at distances ranging from 49 to 290 bp from each other (Figure 2). The PCR amplification products obtained by the application of these primer pairs to human genomic DNA were cloned and sequenced. The comparison of the genomic with the cDNA sequences revealed the positions of the exon-intron borders in the human ACAT2 gene (Figure 2). The gene consists of at least seven exons interrupted by six introns; the number could be higher should the 5' - and 3' -UTRs turn out to contain additional introns. Because the ACAT2 gene is conserved, it can be assumed that it has a similar exon-intron organization in other vertebrates.

Strategy: Before the start of this study, the ACAT2-TCP1 gene overlap had been known to exist only in representative species of the mammalian orders Primates (humans, Ashworth 1993) and Rodentia (house mouse, Dudley et al. 1991). Because primates and rodents were among the first orders to diverge from each other during the adaptive radiation of modern eutherian mammals (Novacek 1992), it can be assumed that the overlap was established before the emergence of extant Eutheria. To trace its evolutionary origin, we therefore turned to representatives of two noneutherian orders, Marsupialia (opossum and wallaby) and Monotremata (duck-billed platypus), as well as representatives of other classes of jawed vertebrates, the Reptilia (cainman), Amphibia (clawed toad), and Osteichthyes/Actinopterygii (zebrafish). In each instance, we first cloned and sequenced the 3' part of the two genes (including a segment of the translated region) to make sure that we identified the homologs of the human/mouse ACAT2 and TCP1 genes rather than some other members of the two gene families. We then carried out Southern blot analysis of genomic DNA from the individual species, using the identified gene segment as a probe to determine whether multiple copies of the gene were present in the genome. Gene pairs thus identified were then tested for an overlap of their 3' ends by PCR amplification with primers complementary to a sequence stretch of the last exon in each of the two genes, using genomic DNA as a template. A failure to amplify a product, even under long-PCR conditions, which should allow an amplification with primers up to 15 kb apart, was taken as evidence that ACAT2 and TCP1 homologs were not overlapping in the species under investigation. Description of the results obtained in the study of the individual taxa follows.

Zebrafish homologs: The zebrafish TCP1 homolog...

aligned using the CLUSTAL W program (Thompson et al. 1994).
was cloned, and the entire coding sequence was determined in a separate study that had been initiated for a different purpose (K. Takami, F. Figueroa, and J. Klein, unpublished results; see Figures 3 and 4). The entire zebrafish ACAT2 coding sequence was determined in two steps using the liver cDNA library. In the first step, the Acat 7 primer spanning the exon 3/4 border was used in conjunction with the vector primer to amplify and sequence the 5′ half of the coding region. (Here we assumed that the exon-intron organization of the zebrafish ACAT2 gene is the same as that of the human gene. The Acat 7 primer was based on the human sequence, which we correctly assumed to be very similar to the zebrafish sequence in this part of the gene.) In the second step, the primer Acat F4, which was based on the zebrafish sequence obtained in the first step, was used in combination with another vector primer to amplify and sequence the 3′ half of the coding region. (Comparison of the zebrafish ACAT2 and TCP1 sequences failed to identify a homology region that would be indicative of an overlap between the two genes. Similarly, both standard and long-PCR experiments using the Acat F8 and Tcp F4 primers failed to yield a product that would be expected

\[
\text{ATG AAT GCA GCC TCA GAT CCT GTC ATC GTC TCG GCG GCG ACC ATC AAT GGT TCC}
\]

\[
\text{TTT AAT GGT GCC TTA GCT GCT TCT GTC AGC CAG CTG GGC TCC ACT GTC ATC AAA GAA}
\]

\[
\text{GTC TTT AAG AOG GCC ACT GTC GCT CCG GAA GAT TGT TGT GAC ACT ATC TTT GGA CAT GTC}
\]

\[
\text{TGT GCA GCA GGC TGT GGG GAA AAT CTT GTT AGACAA GCC AGT GTG GGT GCA GGA GGC ATG GAA}
\]

\[
\text{TAC TCT GTT CCA GCA AGG AGC TGC CAG ATG ATC TGT GGG TCA GGC CTA AAA GCT GTG TGC}
\]

\[
\text{AAT ATG AGC AAG GCT CCT CAC TTG GCT TAC TTG AGA AGA GGA GTA AGA ATA GTG GAG ATG}
\]

\[
\text{CCA CTG ACT GAC AGT ATA CTC TGT GAT GGT CTT ACA GAT GCA TTT CAC AAC TGT CAT ATG}
\]

\[
\text{GCT ATT ACA GCT GAA AAT GTA GCC ACA AAA TGG CAA GTG AGT AGA GAA GCC CAG GAC AAG}
\]

\[
\text{GTT GCA GTT CTG TCC CAG AAG AGC AAG AGC ATC CCT CAG AAA GCT GCC CAT TTT GAC AAA}
\]

\[
\text{GAG ATG GTG CCA GTT TGG GTG TCA ACT AGA AAA GGT CTT ATG GAA GTT AAA ACA GAT GAG}
\]

\[
\text{TTT CCT GSC CAT GGG AGC AAC AGG AAT GAA GCC ATG TCC AAG CTA AAG CCT TAC TTT CTT ACT}
\]

\[
\text{GAT GGA ACG GGA ACA GCC ACC CCA GCC AAT GCT TCA GGG ATA AAT GAT GGT GCT GCA GCT}
\]

\[
\text{GTT CCT CTT ATG AAG AAG TCA GAA GCT GAT AAA GTG GGG CTT ACA CCT TTA GCA CGG ATA}
\]

\[
\text{GTT TCC TGG TCC CAA GTG GGT TGT GAG CCT TCT ATT ATG GGA ATA GGA CCA ATT CCA GCC}
\]

\[
\text{ATA AAG CAA GCT GTT ACA AAA GCA GGT TGG TCA CTG GAA GAT GTT GAC ATA TTT GAA ATC}
\]

\[
\text{AAT GAA GCC TTT GCA GCT GTC TCT GCT GCC ATC GTT AAA GAA CTT GGA TTA AAC CCA GAG}
\]

\[
\text{AAG GTC AAT ATT GAA GGA GGG GCT ATC GCC TGG GCC AC CCT GTT GGA GCA TCT GCC TGT}
\]

\[
\text{GCC CATG ATG GGT GGG ATG GGA ATA GCA ATG TGT GTT CAG AGA GAA TGA}
\]
**TCP1**

**ACAT2**

---

**Figure 3.** Amino acid sequence of zebrafish, clawed toad, and caiman TCP1 and ACAT2 deduced from the corresponding cDNA sequences. Simple majority consensus is shown at the top. Identity with the consensus is indicated by dashes (–), absence of information by dots (•), and gaps introduced to optimize the alignment by asterisks (*). The ▽ and # symbols indicate the beginning of a frame shift and premature stop codons, respectively. Amino acids are given in the single-letter international code.
Overlapping Genes

Zebrafish

**ACAT2**

```
...GCTATGTCATGAGAGAGATGA
```

**TCP1**

```
...GCTCGTTCGAGCTGAGAATAG
```

African clawed toad

**ACAT2**

```
...GCTATGTCATGAGAGAGATGA
```

**TCP1**

```
...GCTCGTTCGAGCTGAGAATAG
```

Caiman

**ACAT2**

```
...GCTATGTCATGAGAGAGATGA
```

**TCP1**

```
...GCTCGTTCGAGCTGAGAATAG
```

Clawed toad homologs: To obtain the Xenopus homolog of the mammalian TCP1 gene, the cDNA library prepared from the jaws of adult frogs (Toyosawa et al. 1998) was PCR amplified using Tcp G8 (a degenerate sense primer based on the comparison of available sequences) in combination with the vector primer. The amplification product of ~469 bp covered enough of the translated region sequence to identify it as the TCP1 homolog. Sequencing of the multiple clones obtained

if the two genes were overlapping. Finally, screening of the zebrafish PAC library by hybridization with zebrafish ACAT2 and TCP1 coding sequence probes and testing of the positive clones by PCR amplification with specific primers revealed the ACAT2 and TCP1 genes to be located on different PAC clones (not shown). Taken together, these three pieces of evidence indicate that the ACAT2 and TCP1 genes are not closely linked in the zebrafish and, hence, not overlapping.
from the single amplification band revealed, however, the existence of two different genes in the Xenopus, TCP1A and TCP1B (Figures 3 and 4). The TCP1B gene has an insertion of 4 bp in exon 7 that is responsible for a frame shift in the rest of the translated sequence (Figure 3). We assume, therefore, that TCP1B is a pseudogene and that TCP1A, which has no identifiable defect in the part sequenced, is the toad’s functional gene. The two genes may be located at distinct loci, possibly in different chromosomes as a result of the genomic tetraploidization that X. laevis is believed to have undergone in its evolutionary history (Ko bel and Du Pasquier 1986). To obtain the Xenopus ACAT2 homolog, we first PCR amplified a product from the cDNA library using the Acat 4 antisense human sequence-based primer in conjunction with the vector primer. In the second step, we used the Acat X1 primer on the basis of the sequence of the product from the first step, in conjunction with the vector primer, to obtain by PCR the 3′-UTR of the clone. The amplification yielded a 750-bp-long fragment containing a large portion of the translated region. Sequencing of several clones isolated from the 908-bp band again revealed the existence of two distinct genes, ACAT2A and ACAT2B (Figure 3). The former appears to be an intact gene, whereas the latter is apparently a pseudogene on account of the presence of at least three premature stop codons (one in exon 2 and two in exon 3) in its sequence (Figure 3). The assumptions made above about the two copies of the TCP1 gene, therefore, also apply to the two copies of the ACAT2 gene. To test the possible overlap between the toad TCP1 and ACAT2 genes, we obtained two primers specific for different sequence stretches of each of the four genes and used them in standard and long PCR in all possible pairwise combinations, always matching a sense ACAT2 primer with an antisense TCP1 primer or vice versa. In none of these combinations did the PCR yield a detectable band. We conclude, therefore, that in the clawed toad, as in the zebrafish, neither of the two ACAT2 genes is overlapping with either of the two TCP1 genes.

Caiman homologs: Reptilian TCP1 and ACAT2 sequences were cloned from a cDNA library prepared from the jaws of a 3-day-old caiman (Toyosawa et al. 1999). The TCP1 clones were obtained from the library by PCR amplification using the Tcp G8 sense and the vector antisense primers. The cloned and sequenced amplification product was 581 bp long and contained a large part of the translated TCP1 region as well as the entire 3′-UTR (Figures 3 and 4). An ACAT2 clone obtained in a similar manner using the Acat X1 (Xenopus-specific) primer was 780 bp long and also encompassed both the 3′-translated and -untranslated regions. Efforts to PCR amplify a product with one primer (Acat C1) located in the translated region of caiman ACAT2 and another primer (Tcp C1) located in the translated region of caiman TCP1 failed, leading us to the conclusion that the two genes are nonoverlapping in this species.

Platypus homologs: Because a platypus cDNA library was not available to us, we resorted to the use of genomic DNA. Under the assumption that the TCP1 and ACAT2 genes either overlap or are closely linked, we used the Tcp G3 and Acat G1 combination of primers annealing to the translated sequences of the corresponding eutherian genes. The PCR amplification of the genomic DNA with these two primers yielded a 493-bp product that upon cloning and sequencing proved to represent the overlapping ACAT2 and TCP1 sequences (Figure 5). The platypus sequence from the ACAT2 stop codon on one DNA strand to the TCP1 stop codon on the complementary strand encompasses 253 bp compared to 215 and 183 bp of the human and mouse sequences, respectively. The unavailability of platypus cDNA precluded a definitive identification of the polyadenylation signals. A putative ACAT2 polyadenylation signal is, however, present at a distance of 237 bp from the ACAT2 stop codon and at a distance of 11 bp from the TCP1 stop codon. Assuming that the polyadenylation site is >15 bp downstream of the signal, the cleavage site of the ACAT2 transcript probably overlaps with the TCP1 translated region on the complementary DNA strand. Similarly, a putative polyadenylation is present on the TCP1 coding strand 87 bp downstream of the TCP1 stop codon and, therefore, the TCP1 transcript cleavage site probably does not overlap with the translated region of the ACAT2 gene on this strand.

Opossum and wallaby homologs: In our study, marsupials were represented by two species, the opossum and the wallaby. A cDNA library, however, was available to us only from the former species. Sequencing of PCR products obtained from both genomic DNA and cDNA (primers Acat G3 and Tcp G1) revealed an overlap of the ACAT2 and TCP1 genes in the opossum, and sequencing of a product of genomic DNA amplification similarly revealed an overlap of the two genes in the wallaby. In the opossum, the distance between the stop codons of the two genes was 266 bp (Figures 1 and 5).

Figure 5.—Nucleotide sequence alignment of a portion of the last exon (boxed) and the overlapping 3′ untranslated regions of the ACAT2 and TCP1 genes in various mammalian species. The human and mouse sequences are from Song et al. (1994), Kubota et al. (1994), and Dudley et al. (1991); the remaining sequences are from the present study. The known extent of 3′ untranslated regions is indicated by arrows. The conserved segment is underlined, and putative polyadenylation signals are indicated by an overhead line. (−−) Identity with the simple majority consensus sequence at the top of the alignment, (±) gaps introduced to optimize the alignment, (−) nonavailability of sequence information.
The ACAT2 polyadenylation signal was located 254 bp downstream from the gene’s stop codon and 6 bp from the TCP1 stop codon on the complementary strand, so the transcript cleavage site overlapped with the TCP1 translated sequence. The TCP1 polyadenylation signal was located 113 bp downstream from the stop codon and at a distance of 147 bp from the ACAT2 stop codon on the complementary strand (Figures 1 and 5). The transcript cleavage site was located 16 bp downstream of the polyadenylation signal and, hence, did not overlap with the ACAT2 translated signal. In the wallaby, the arrangement of the polyadenylation signals of the ACAT2 and TCP1 genes was similar to that in the opossum, except that the distance between the two signal sites was somewhat longer (116 bp in the wallaby compared to 101 bp in the opossum).

**DISCUSSION**

The results of the present study suggest that for most of vertebrate evolutionary history, the ACAT2 and TCP1 genes have been independent entities, as indeed they still are in modern bony fish and presumably all tetrapods except mammals. In the extant nonmammalian gnathostomes, the two genes are either not linked or, if they are on the same chromosome, they are at a distance from each other that precludes the formation of an overlap between them. The overlap observed in all mammals tested, including the monotremes and the marsupials, may have therefore arisen during the transition from therapsid reptiles to mammals (Carroll 1998), presumably by a chromosomal rearrangement that brought the two genes together. The arrangement may have been a translocation, a deletion if the ACAT2 and TCP1 genes had already been located on the same chromosome, or a more complex event that may have involved several successive steps. The overlap may have arisen in one of two ways. First, the rearrangement may have been accompanied by the loss of a part of the 3′-UTR, including the polyadenylation signal from, say, the TCP1 gene. By chance, however, the 3′-UTR of the new neighbor, the ACAT2 gene, contained on the noncoding strand all the signals necessary for the termination of transcription and processing of the transcript so that the TCP1 gene could continue to function normally. Second and perhaps more likely, the two genes became neighbors through the rearrangement but at first did not overlap. Only later, when one of the genes lost its original polyadenylation signal and began to use a signal that happened to be present on the noncoding strand of the other gene, did the pair become locked in. The noncoding strands of the ACAT2 genes in the zebrafish, toad, and caiman do indeed contain one or more potential, correctly oriented and spaced polyadenylation signals in their 3′-UTRs that could be used by the TCP1 gene if the two genes were to come together now (Figure 4). The abundance of potential polyadenylation signals in the noncoding strand is undoubtedly caused by the relatively high AT content (60–70%) that characterizes the 3′-UTRs of the ACAT2 and TCP1 genes.

The rearrangement thus generated a genetic odd couple that has henceforth been inherited as a unit. The restriction of the overlap to a single phylogenetic lineage, the mammals, suggests that the link-up of the two genes occurred only once and then persisted for more than 200 million years. This conclusion is further supported by the observation of sequence conservation in the overlapping 3′-UTRs of the ACAT2 and TCP1 genes (Figure 5). The sequence similarity among the mammalian sequences is poor in the part of the overlap flanking the ACAT2 translated region, but is rather striking in the part flanking the TCP1 translated region, in which a whole sequence block has been conserved during evolution of monotremes, marsupials, and eutherians from their common ancestors (Figure 5). For 53 bp of the alignment up to the ACAT2 stop codon, the comparisons of human with mouse, opossum, wallaby, and platypus sequences, respectively, give 71, 78, 76, and 71% identity, with only seven indel events postulated for alignment. The presence of the postulated ACAT2 polyadenylation signal in this segment may account for some of the conservation, although the extent of conservation appears to exceed simple polyadenylation signal requirements. To examine whether such conservation is common in eutherian to noneutherian comparisons, we have collated the 3′-UTRs of 10 sequences available from eutherian and metatherian sources and selected at random from DNA databases (spem protein Sp17, α-tumor necrosis factor, preprolactin, occludin, transthyretin, β-casein, preprorouguanylin, pyruvate dehydrogenase E1-α, β-actin, and protamine P1). We examined the sequences by dot plot, using sensitivities of 70% matches in windows of 10-bp size, and we assessed the degree of homology. Although crude, this method avoids the problems of changes in 3′-UTR length, gap penalties, and base composition corrections in the alignment of such poorly conserved sequences. Among the 10 sequences selected, only 1 (β-actin) exceeds the homology found in TCP1-ACAT2 3′-UTR comparisons, and an additional 3 (α-tumor necrosis factor, transthyretin, and protamine P1) show comparable degrees of conservation. The other 6 sequences show little or no sequence similarity between metatherian and eutherian 3′-UTRs.

Human-mouse comparisons show that the ACAT2 and TCP1 genes evolve at moderately high synonymous substitution rates that are comparable to the average rates observed in large-scale surveys of mammalian genes (O’Huigin and Li 1992). Hence, relative to the synonymous rate of the rest of the ACAT2 and TCP1 genes, the rate has slowed in parts of the overlap region. We assume, therefore, that the sequence similarity observed in the overlap region is retained by selection for reasons...
related to the overlap. Miyata and Yasunaga (1978) have argued that the rate of evolution can be expected to slow down in the overlapping stretches, but their argument applies to the overlap of translated regions only; they reason that the proportion of nondegenerate sites is higher in overlapping genes than in nonoverlapping ones, thus reducing the proportion of synonymous substitutions relative to the total number of substitutions. In the 3'-UTR, the slowdown in overlapping stretches might be related to the retention of signals or tertiary structures necessary for processing of the transcript. The overlap might restrict the number of permissible substitutions in certain parts of the 3'-UTR.

Finally, the conclusion that the observed overlaps derive from the same event is also supported by the fact that the relative arrangement of the polyadenylation signals is very similar in the platypus, wallaby, opossum, and human (Figure 1). Only in the mouse does the ACAT2 3'-UTR appear to extend all the way into the last intron of the TCP1 gene (Ashworth 1993). This use of an alternative ACAT2 polyadenylation signal appears to predate the duplication of the gene because both ACAT2 copies have a 3'-UTR extending into the last intron of TCP1. In all other known mammalian sequences, the distances of the polyadenylation signals from the stop codons are comparable.

The reasons the ACAT2 and TCP1 genes have remained coupled together for some 200 million years are unclear. In fact, we cannot exclude the possibility that in some of the 4629 living species of mammals (Wilson and Reeder 1993) they have not. But even if a secondary separation has occurred in some species, it is probably safe to assume that in the great majority of mammals, the two genes have stayed together. There is no evidence that ACAT2 and TCP1 are in any way related to each other evolutionarily, structurally, or functionally. The same is true for the other documented cases of gene overlap in vertebrates (Williams and Fied 1986; Adelman et al. 1987; Emi et al. 1988; Morel et al. 1989; Cawthon et al. 1991; Shayig and Avadhani 1992 and other references cited earlier). It is also difficult to imagine any potential advantages that the gene overlap might have for the participating loci. In fact, it may have certain disadvantages manifested when the couple duplicates and then begins to undergo cycles of expansions and contractions, which has been well documented especially for the mammalian CYP21-C4 pair (Morel et al. 1989; Kawaguchi et al. 1991). In such cases, deletions and other more complex changes may lead to deficiencies with clinical consequences, which in this case would affect two genes instead of one, as they might in the case of uncoupled genes. In the absence of any known advantage, the persistence of gene overlap may be a consequence either of the conservative nature of the evolutionary process or of the difficulties associated with the separation. A divorce of the two genes would require that one or both of them would find, after their separation, all the signals necessary for the termination of transcription and the processing of the transcript in their immediate vicinity. The signals would have to be on the right strand, in the right order, and at the appropriate distances from one another. The probability of this happening might be quite low.

We thank Ms. Jane Kraushaar for editorial assistance, as well as Prof. W. H. Stone and Dr. R. W. Slade for tissue samples.

LITERATURE CITED


Gilbert, D. G., 1995 SeqPup Version 0.4j; a biosequence editor and analysis application, http://libubio.indiana.edu/soft/molbio/.


Williams, T., and M. Fried, 1986 A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3′ ends. Nature 322: 275–279.


Communicating editor: N. Takahata