Schonbaum, G. R., Zerner, B., & Bender, M. L. (1961) *J. Biol. Chem. 236*, 2930–2935.
Sugiura, M. (1984) in *Lipases* (Borgstrom, B., & Brockman, H., Eds.) pp 505–524, Elsevier, Amsterdam.
Vainio, P., Virtanen, J. A., & Kinnunen, P. K. J. (1982) *Biochim. Biophys. Acta 711*, 386–390.
Walsh, C. (1979) *Enzymatic Reaction Mechanisms*, pp 84–86, Freeman, San Francisco.

Structure of the Human Thyroid Peroxidase Gene: Comparison and Relationship to the Human Myeloperoxidase Gene†

Shioko Kimura,*,‡ Young-Sook Hong,‡,§ Tomio Kotani,‖ Sachiya Ohtaki,‖ and Fumitaka Kikkawa‡
Laboratory of Molecular Carcinogenesis, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, and Central Laboratory of Clinical Investigation, Miyazaki Medical College Hospital, Kiyotake, Miyazaki, 889-16, Japan
Received October 31, 1988; Revised Manuscript Received January 26, 1989

ABSTRACT: All exons of the human thyroid peroxidase gene were cloned from phage and cosmid libraries and sequenced, including 2599 base pairs of upstream DNA. The gene contains 17 exons and covers at least 150 kilobase pairs of chromosome 2. The transcription start site was identified by both S1 mapping and primer extension; a typical TATA box was found 25 bases upstream of the putative start site. A comparison of the gene structures of thyroid peroxidase and a granulocyte protein, myeloperoxidase, revealed that the positions of the 3rd through 11th exon–intron junctions in thyroid peroxidase coincide exactly with those of the 2nd through 11th exon–intron junctions in myeloperoxidase except the 7th myeloperoxidase junction, that does not have any counterpart in thyroid peroxidase. The amino acid codon separation pattern in each junction is well conserved between both enzymes. Four exons, unique to thyroid peroxidase, are located at the 3′ end of the gene (exons 13–16), each of which encompasses a different protein module. Three of these modules, representing exons 13, 14, and 15, bear significant similarities to C4b-β2 glycoprotein, the EGF-LDL receptor, and a typical transmembrane domain, respectively. The genes coding for these modules were probably fused to an ancestral peroxidase gene to generate the present thyroid peroxidase gene. The data suggest that intron loss, and/or insertion, and exon shuffling have played important roles in the evolution of the thyroid peroxidase gene.

Thyroid peroxidase (TPO;[1] donor:hydrogen-peroxide oxidoreductase, EC 1.11.1.7) is a membrane-bound hemoprotein which is involved in the biosynthesis of thyroid hormones. It catalyzes both the iodination and the coupling of tyrosines in thyroglobulin to yield thyroxine (T4) and triiodothyronine (T3) (DeGroot & Niepomniszcze, 1977; Nunez, 1980). Human TPO was purified by monoclonal antibody assisted chromatography and appeared to have two bands in the 107-kilodalton region of an SDS–polyacrylamide gel (Czarnocka et al., 1985; Ohtaki et al., 1986). TPO has been shown to be a major component of the thyroid microsomal antigen involved in autoimmune thyroid diseases (Czarnocka et al., 1985; Portmann et al., 1985; Kotani et al., 1986). Recently, the cDNA clone to human TPO has been isolated, and its nucleotide and deduced amino acid sequences have been determined (Kimura et al., 1987; Libert et al., 1987a,b; Magnusson et al., 1987). The evidence for the identity of TPO as the microsomal antigen expressed in patients with autoimmune thyroid diseases was further provided by cDNA cloning (Libert et al., 1987b).

A striking homology was found between human TPO and human myeloperoxidase (MPO; Johnson et al., 1987; Morishita et al., 1987a), indicating that both enzymes are members of the same gene family and evolved from a common ancestral gene (Libert et al., 1987b; Kimura & Ikeda-Saito, 1988). Interestingly, the latter enzyme is expressed in granulocytes and monocytes and plays a major role in the H_2O_2-dependent microbicidal system of neutrophils. The tissue-specific expression patterns and physiological roles of TPO and MPO are quite distinct. Due to the similarities of their cDNAs, we sought to examine the relationships between the structure of both peroxidase genes.

The MPO gene has recently been isolated and sequenced (Morishita et al., 1987b). This gene is about 10 kilobase pairs (kbp) long, contains 12 exons (Morishita et al., 1987b), and is located on human chromosome 17 (Weil et al., 1987; Liang et al., 1987). On the other hand, the human TPO gene has been mapped to chromosome 2 (Kimura et al., 1987; Libert et al., 1987b). In the present report, we describe the cloning and complete exon structure of the human TPO gene. The gene spans over 150 kbp and consists of 17 exons and 16 introns. A comparison of exon–intron junctions between the TPO and MPO (Morishita et al., 1987b) genes revealed that the position of the junctions and the amino acid codon separation patterns of both enzymes are well conserved. In addition, the 3′ portion of the TPO gene has the insertion of four unique gene modules which are not observed in the MPO gene. The possible evolutionary events responsible for the divergences and the conservations of DNA sequence between TPO and MPO genes will be discussed.

Materials and Methods

Cloning and Sequencing of the Human TPO Gene. Total DNA was isolated from a blood donor's lymphocytes by a

† The nucleic acid sequence in this paper has been submitted to GenBank under Accession Number J02856.
‡ National Cancer Institute.
§ Present address: College of Medicine, Ewha Womans University, Seoul, Korea.
‖ Miyazaki Medical College Hospital.

[1] Abbreviations: bp, base pair(s); kbp, kilobase pair(s); TPO, thyroid peroxidase; MPO, myeloperoxidase.

0006-2960/89/0428-4481$01.50/0

standard protocol. The cells were lysed by osmotic shock and treated with proteinase K and RNase. DNA was then purified by successive phenol extractions, taking care to avoid shearing effects. The DNA was dialyzed exhaustively against 10 mM Tris, pH 8.0, and 1 mM EDTA. This DNA was partially digested with *Mbo*I for various length of time and fractionated by sedimentation on either a 5–25% or a 5–30% NaCl gradient (in 10 mM Tris-HCl, pH 7.5, and 1 mM EDTA), and fractions containing fragments of 20–25 or 35–50 kbp were ligated to either λEMBL3 or Sp cos 2 (Heilig et al., 1987), respectively. The libraries were screened without prior amplification by plaque or colony hybridization, respectively, using the human TPO cDNA, previously designated as phTPO-1 (Kimura et al., 1987). An amplified human genomic library, constructed in the vector Charon 4A by partial *Hae*III/*Alu*I digestion of human DNA (Maniatis, 1978), was also screened, and two clones covering exons 3, 8, 9, and 10 (Figures 1 and 3) were obtained. Restriction fragments derived from the genomic clones that hybridized to the cDNA were subcloned into pUC9. These fragments were used to make M13 shotgun libraries (Deininger, 1983), and M13 clones that gave positive signals after hybridization to the cDNA were chosen for sequencing. Sequencing was carried out by using the dideoxy nucleotide chain termination method (Sanger et al., 1977), and sequence data were compiled and analyzed by the Beckman Microgenie sequence analysis program.

For mapping of λEMBL3 or cosmid clones, individual clones were partially digested with *Eco*RI, and the digests were analyzed by electrophoresis on 0.4% and 1.2% agarose gels, Southern blotting, and hybridization to the TPO cDNA and cos site specific oligonucleotides (New England Biolabs, Beverly, MA; Rackwitz et al., 1984). In order to precisely determine the exon sites, each clone was also treated with 14 different restriction enzymes, including *Eco*RI, by single, double, or triple digestion, followed by Southern blotting and hybridization to the TPO cDNA.

Determination of the Transcription Start Site. The start site of transcription was determined by both S1 nuclease mapping and primer extension analysis using 5′ end-labeled specific oligonucleotides (positions −17 to +43 and +24 to +43, respectively; Figures 2 and 3) as described (Ausubel et al., 1987). The human thyroid poly(A) RNA used in this study was obtained as previously described (Kimura et al., 1987). Oligonucleotides were synthesized on an Applied Biosystems Model 380B DNA synthesizer (Foster City, CA). Products of the S1 nuclease protection and primer extension reactions were electrophoresed on 8% polyacrylamide–50% urea gels.

Results and Discussion

Isolation of the Human TPO Gene. We previously isolated two human TPO cDNA clones, phTPO-1 [3048 base pairs (bp)] and phTPO-2 (2877 bp). The phTPO-2 is 171 nucleotides shorter than phTPO-1 due to the alternative splicing of the pre-mRNA in the middle of the coding sequence (Kimura et al., 1987). Using the hTPO-1 cDNA, a human genomic library, constructed in λEMBL3, was screened, and several clones were isolated. These were grouped into different classes on the basis of restriction enzyme fragmentation patterns. *Eco*RI fragments derived from representatives of each group of clones that hybridized to the cDNA were then subcloned into pUC9. In some cases, due to a fusion of human DNA to the junction of the phage arms, *Eco*RI/*Sal*I-digested fragments were used for subcloning. This *Sal*I site is derived from a cloning site in the λEMBL3 vector, and, therefore, the subcloned *Eco*RI/*Sal*I fragments do not represent authentic *Eco*RI/*Sal*I genomic fragments. Restriction enzymes *Bam*HI and *Bgl*II were also used to isolate fragments from phage clones when these enzymes gave shorter and therefore more convenient-sized fragments than *Eco*RI for subcloning and sequencing. Each fragment thus obtained was subjected to M13 shotgun sequencing as described under Materials and Methods. After partial sequencing, none of the fragments were found to contain cDNA sequence which later turned out to contain exons 3, 8, 9, 10, and 16 (Figures 1 and 3). Further screening of the same λEMBL3 library did not result in the isolation of any clones containing these exon sequences. However, screening of an amplified human genomic library in the vector Charon 4A (Maniatis, 1978) yielded two new clones, one of which contains exon 3 and the other of which contains exons 8, 9, and 10. The cloned DNA containing exon 16 was not found in any of the phage libraries even after repeated screening. This is likely to be caused by the presence of sequence around the 16th exon which inhibits growth of recombinant phage in *Escherichia coli*. This sequence may be a relatively short sequence (200–500 bp) that contains inverted repetitions, which form snap-back and stem-and-loop structures (Leach & Stahl, 1983; Wyman et al., 1985). It has been shown that a large proportion (8.9%) of the phage carrying different fragments of the human genome fail to grow on commonly used *rec*$^+$ hosts due to such secondary structures (Wyman et al., 1985). A failure to obtain phage clones containing certain portions of the sequence has also been reported in human thyroglobulin gene (Baas et al., 1986). We then constructed a cosmid library to search for a clone containing exon 16 sequence and to map the TPO gene. A clone carrying the exon 16 sequence was successfully isolated from the cosmid library.

In Southern blot analysis of *Eco*RI-digested genomic DNA, the approximate cumulative length of all *Eco*RI fragments that hybridized with full-length hTPO-1 cDNA was about 95 kbp (data not presented). Most of the *Eco*RI fragments found in the original phage clones corresponded in size to these genomic *Eco*RI fragments. The fragment containing exon 10, however, did not correspond in size to any fragment detected on the Southern blot, suggesting that deletions or recombinations might have occurred during amplification of the Charon 4A library. Alternatively, this could be due to a genomic restriction fragment length polymorphism (RFLP) since the DNA used for the Southern blot experiments is different from that used for the original construction of the Charon 4A library (Maniatis, 1978). Two *Eco*RI/*Sal*I subcloned fragments, that were artificially created by the *Sal*I site in the vector λEMBL3, contained 11th and 15th exon sequences. By hybridization of Southern genomic blots with partial cDNA fragments containing sequences specific to either of these exons, the 11th and 15th exons were found to reside within 23 and 5.5 kbp *Eco*RI fragments, respectively. On the basis of results of sequencing genomic clones, genomic Southern blotting, and restriction mapping of phage and cosmid clones, a compiled map of the human TPO gene can be drawn (Figure 1). The total length of the TPO gene is currently not known, but we believe it is at least 150 kbp. These genomic cloning results also confirmed the previous prediction that the human TPO gene is likely to be a large single gene (Kimura et al., 1987).

Initiation Site for Transcription. To determine the transcription start site in the TPO gene, primer extension and S1 nuclease analysis were performed (Figure 2). Both revealed two clusters of start sites, each of which consists of two to four bands. The size of the longest extended fragment obtained after primer extension analysis was one base different from

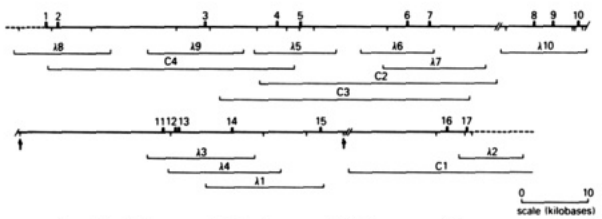


FIGURE 1: Partial map of the human TPO gene. The composite TPO gene is indicated by the *solid bar*. The positions of exons are indicated by the *black boxes*. *Eco*RI sites used for subcloning and sequencing are indicated by small vertical bars except for *Eco*RI sites denoted by arrows that were estimated on the basis of the size of *Eco*RI fragments predicted from genomic Southern blots, hybridized with DNA sequence specific to exon 11 or 15. The positions of the individual lambda (λ) and cosmid (C) clones are shown under the composite gene.

that of the S1 nuclease protected fragment. This may be due to the presence of an mRNA cap structure that prohibits complete primer extension. Moreover, the sizes of the fragments in the lower cluster in primer-extended products and S1-protected fragments were also not identical. The reasons for these slight discrepancies are not clear, but these results suggest that two sets of transcription start sites exist in the TPO gene. The start site was assigned to the longest primer-extended band.

Sequence of the Human TPO Gene. The nucleotide sequences of all exons and their flanking regions of the human TPO gene including 2599 bp upstream, a total of about 32 kbp, were determined (Figure 3). The gene consists of 17 exons and 16 introns. All of the intron splice donor and acceptor sites follow the GT---AG rule (Mount, 1982; Figure 3 and Table I). The sequence 5′-TATAA-3′ is found 25 nucleotides upstream of the putative cap site. The protein initiation codon ATG, following a putative signal peptide, and some of the amino-terminal residues of the mature protein (Kimura et al., 1987; Libert et al., 1987b; Magnusson et al., 1987; Kimura & Ikeda-Saito, 1988) are observed in the second exon. The 171-nucleotide sequence that was absent in the previously reported 2nd cDNA clone, phTPO-2 (Kimura et al., 1987), is contained in the 10th exon. Recently, the positions of the putative proximal histidine residue that binds to the iron center of the enzyme, and the two possible distal histidine residues, one of which is important for peroxidase activity together with the nearby arginine residue, have been reported (Kimura & Ikeda-Saito, 1988). These histidine-containing regions are coded for by the latter half of exon 8 and exons 9 and 10, respectively.

Libert et al. (1987b) have reported that in the C-terminal portion of human TPO, two gene modules exist in precise juxtaposition, one of which belongs to the C4b-β2 glycoprotein gene family (Davie et al., 1986) and the other to the EGF-LDL receptor gene family (Doolittle et al., 1986). These are followed by the transmembrane segment of the protein. The boundaries of these gene modules agree very well with exons 13, 14, and 15, respectively. Furthermore, a sequence similar to the cytochrome *c* oxidase subunit I (Libert et al., 1987b) is located on the latter and the former half of exons 9 and 10, respectively. Of particular interest is that the autoantigenic region proposed by Libert et al. (1987b) is coded entirely by DNA in exon 11.

Comparison of the Structural Organization between Human TPO and MPO Genes. The human TPO (3048 bp; Kimura et al., 1987; Libert et al., 1987a,b; Magnusson et al., 1987) and MPO (3213 bp; Johnson et al., 1987; Morishita et al., 1987a) cDNAs display 46% and 44% [42% in Libert et al. (1987b)] similarities in their nucleotide and deduced amino

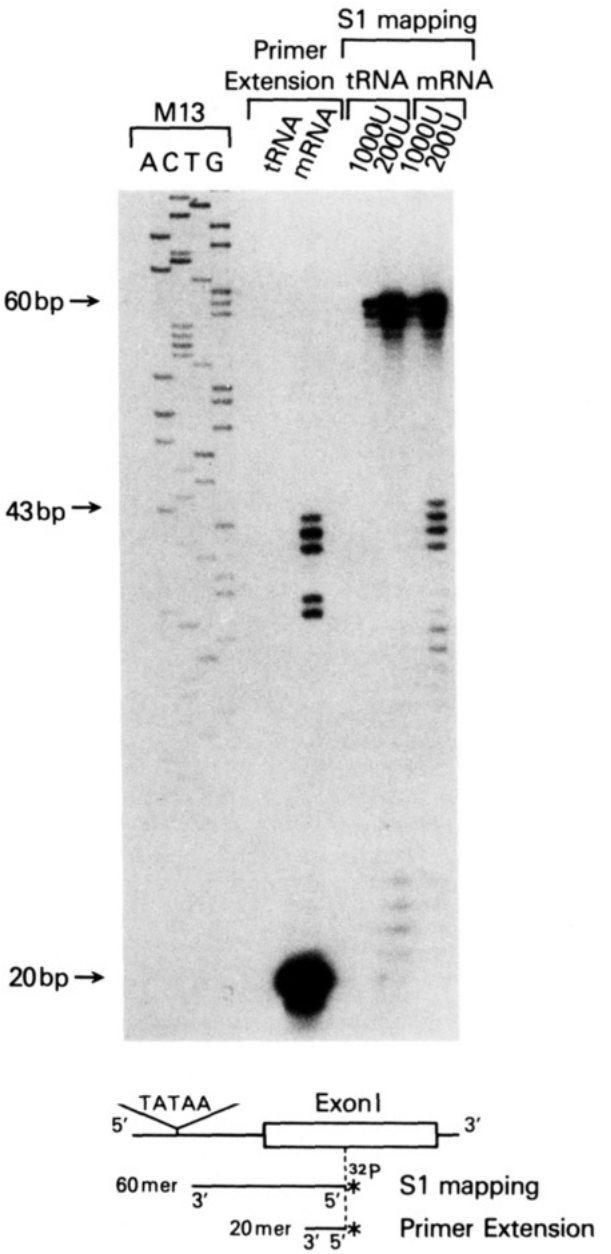
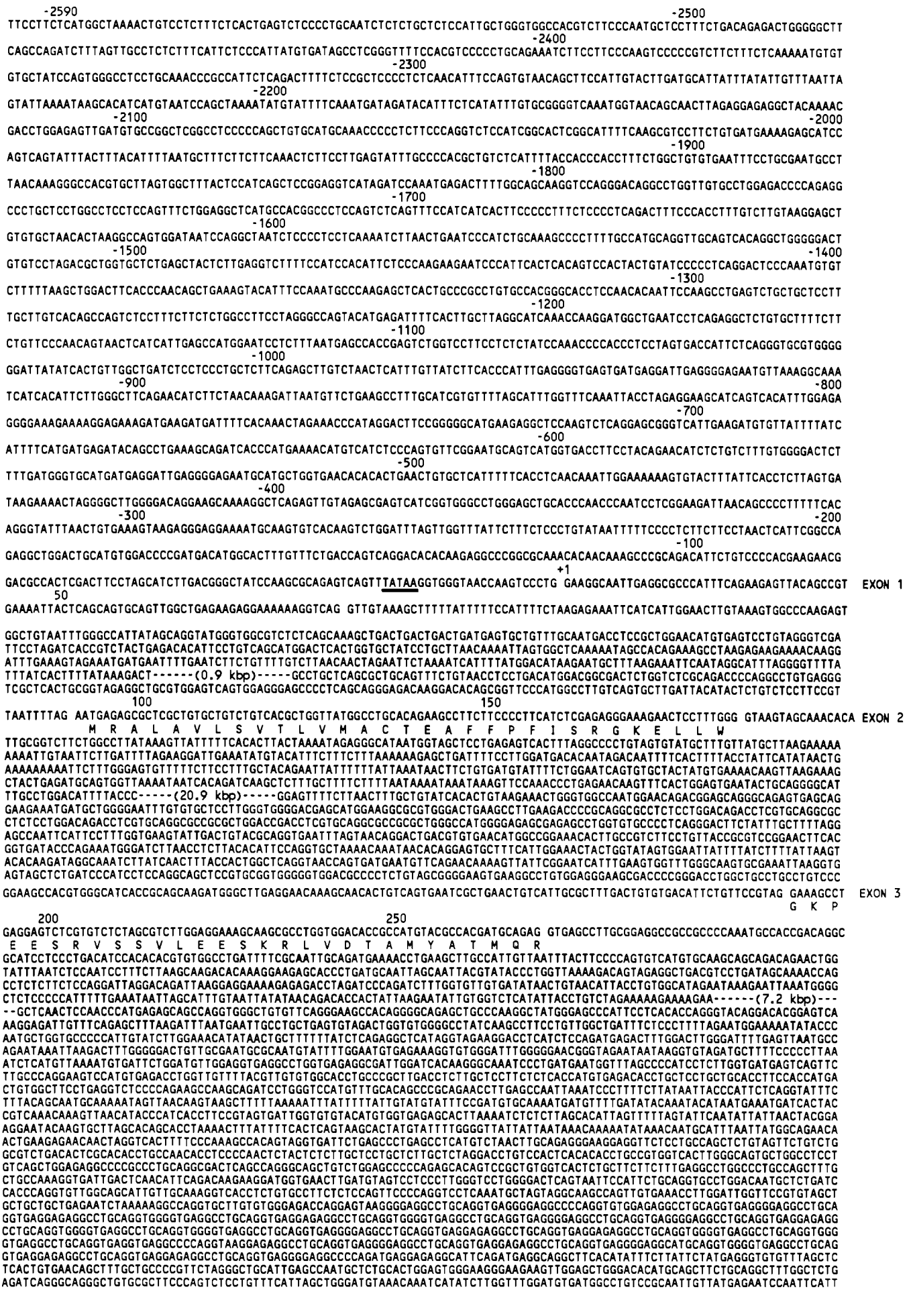
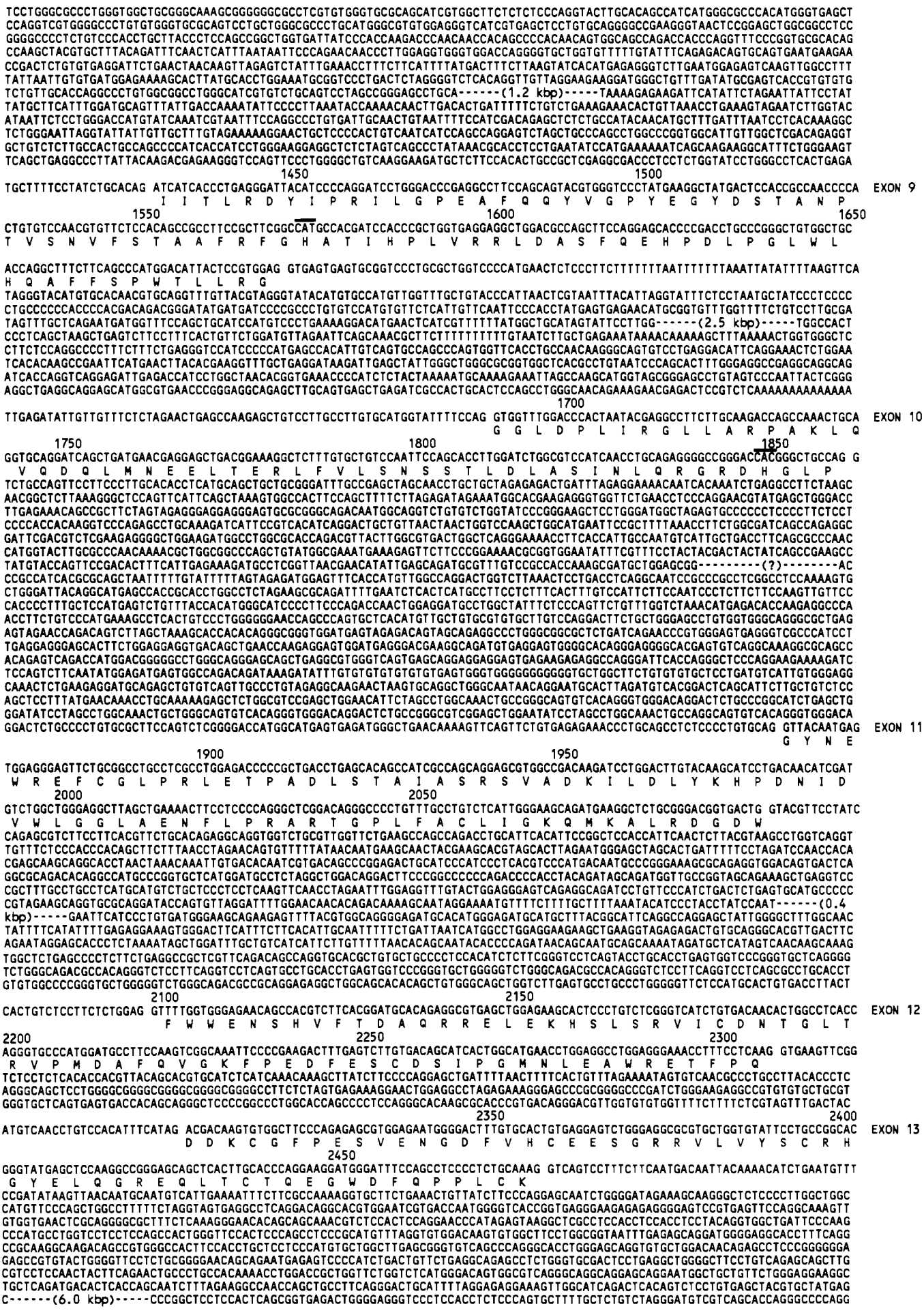


FIGURE 2: Primer extension and S1 nuclease mapping of the human TPO gene. S1-protected fragments generated by two different concentrations of S1 nuclease and the primer-extended fragments were electrophoretically separated on an 8% polyacrylamide–50% urea gel. M13mp18 single-stranded DNA (*four lanes at left*) was sequenced to generate a ladder of fragment-size standards in the same gel. Yeast tRNA was used as a control. The 60-mer and 20-mer oligonucleotides were used as a probe for S1 mapping and as a primer for primer extension, respectively. Fragments at 60 bp represent remnants of undigested S1 probe, and the fragment at 20 bp is unextended primer from the primer extension reaction. The longest fragment generated by primer extension is denoted by the arrow at 43 bp. This was assigned as the start site (position +1) in the TPO gene in Figure 3.

acid sequences, respectively, which clearly indicated that both enzymes evolved from a common ancestral gene (Libert et al., 1987b; Kimura & Ikeda-Saito, 1988). The human MPO gene is located on chromosome 17 (Weil et al., 1987; Liang et al., 1987) and spans about 10 kbp, containing 12 exons and 11 introns (Morishita et al., 1987b), whereas the human TPO gene was mapped to chromosome 2 (Kimura et al., 1987; Libert et al., 1987b), covers at least 150 kbp, and consists of 17 exons and 16 introns (Figures 1 and 3). It was, therefore, interesting to compare the structure of both genes.


-2590
-2500
-2400
-2300
-2200
-2100
-2000
-1900
-1800
-1700
-1600
-1500
-1400
-1300
-1200
-1100
-1000
-900
-800
-700
-600
-500
-400
-300
-200
-100
+1
EXON 1
50
(0.9 kbp)
100
150
EXON 2
M R A L A V L S V T L V M A C T E A F F P F I S R G K E L L W
(20.9 kbp)
EXON 3
G K P
200
250
E E S R V S S V L E E S K R L V D T A M Y A T M Q R

```
CTGATTTGGCAGTTTTGTAAAGGGCAAGTCTGTGCCATTTTCCTCATCATTCAATACTGATGCATCACTTACACCCTATCCTGTCCTAAACCCAGTGACCCCGTGGAACAGCAACTGCAC
ATGTTTTACCTGTGCACACCCCGCAGTGCCTGTCACATTGTCTGGGACACAGTAGTTACTCAATAAATGTCTGCTTAATTAATTAGTAATTAAGTACCAAAGATACCATAGACAAATAAT
                                                                300                                                 350
CTATTTTATATCTTCTTTATGTGCCATAG AAACCTCAAGAAAAGAGGAATCCTTTCTCCAGCTCAGCTTCTGTCTTTTTCCAAACTTCCTGAGCCAACAAGCGGAGTGATTGCCCGAGCA  EXON 4
                                N  L  K  K  R  G  I  L  S  P  A  Q  L  L  S  F  S  K  L  P  E  P  T  S  G  V  I  A  R  A
                                                400
GCAGAGATAATGGAAACATCAATACAAGCGATGAAAAGAAAAGTCAACCTGAAAACTCAACAATCACAGCATCCAACGG GTAATGTGTGCCCCTCTCCCCACTGAGGAGTGGCAACTCCC
 A  E  I  M  E  T  S  I  Q  A  M  K  R  K  V  N  L  K  T  Q  Q  S  Q  H  P  T
GAAGGAGGACACCTTGACTTTGTGCAGGGGCTGCTTGCTCAGGGCATGTTTTTTACTTGAGACCAAGCAGGATGGGACTCCAGCTCTTTCAAAGCATACAAGCTGCAATTTTTAAAAATG
CTTACCAAACAGTCTCAATATCTTATTTTTGTTACTCTGGAAAATAGCCATTTCTCTCCTATCCAAATCAAATTTGAGAAGTAGCTCAATAAGCTAAACATCCATTTTTTATGGTATAAA
ATAATTTAGGAATTTTCCCATAGCCCCCTCTTAGTTAAAACTTGGCAAGCTCATTCATAAAACCTCTACAGTGAAAGGACTTCATGTGTGAGGCAGCGCATTGCAGGGAGCCCCATCTTC
CCACTGAGTTGCTGGTGCCTCAGCAGAACTACCCAGGACATTTGAGAGCCAATTTACTCATCTCTAAAATGAGGACAGTGGCACTGCTCACTCCAATGGCTATCAGAAGACCAGTGCATA
AACCAGGAGCAAATGTCGTCATTTCTTCAGCTCTAAAGCACTCTATGGGCATAAGATACTGAGATGGTAATGCAGGCAGTCAAAAGGCATATTGGTGCTTCCTATGATGCAGAGCAAGTA
AGCGAAATGTCTTTTTAATTAATTTATCAAGAATTTTTTCTCTGCAGCGCCATGAGCTACTTGGCCTCTGAGTTCTTCCTTCTAACTAGCTATAGATCAATTTGGTGCTGGTTACTTAGC
TTTTAGAATAAGACAAACTTGGCTAAATCAGTATTGACCAGAATTCAAGGGGTGGTCTGGCTTCTTCTAACAACCAACCATCAGATATGAGAGCAAAGAAATGATTTGTATTTGGAAATT
                                                                                                                    450
ATATTTAAAGGGAAGCAGAGCATAACATTTGAAATTTGCCAGGCTAGCACTTGATAGGAA------(2.4 kbp)-----GAATTCATGGTTTCCTATTTTTCACAG ATGCTTTATCAGA  EXON 5
                                                                                                                D  A  L  S  E
                                                500                                                 550
AGATCTGCTGAGCATCATTGCAAACATGTCTGGATGTCTCCCTTACATGCTGCCCCCAAAATGCCCAAACACTTGCCTGGCGAACAAATACAGGCCCATCACAGGAGCTTGCAACAACAG
  D  L  L  S  I  I  A  N  M  S  G  C  L  P  Y  M  L  P  P  K  C  P  N  T  C  L  A  N  K  Y  R  P  I  T  G  A  C  N  N  R
GTATTGTTTGTGGATTTTCTTAATCTTCTCGTGAAAGTTGGATCTGAACATGACTAGATGCCATCATGAAGGAACTTCTACACCACTGGTGGATCTGTATCCACCCCTGAGCCCCTGGTT
CCTGTCCTTGGCCTCCCCGACATGGCCTCCTGCATGCCTGGTGGTGCCTCAAGCAGCGTGCCAAGCAGTACTGTCATTCCCAGCCACAACGTCCCATCCACATAGGTGCTTACAGGTGAC
TCTTAACTTCCTAACCCCACAAATTCCACCGCCCTTCACCTAAGAGCTCAAGCCACGTGCTTAGGCGCCATCCTTGATTTCGCTAGCTCCGCTACCCATTGTTTCCAATGCACCCCCAAG
ACTGTCACTTTTACCTCCAAAATGAGTCCTCTCCACTCCCTGTCACCTTAGCTGAAGCCAAAATAACCATCCAGCTGCCATCACGTGGTGGTCCCCTCACTGGTTCTCCTAATTTAACTC
TTGCTTTCCCCATTCACGCTCCCTGTTATGGCCAACACCATCTCAGCACCCTCCTTCTCAACTCCTCCCTCAGTCCCCACTGCACTCAGAGTGAGTTCAGATCCGCCCAACAGTGGTGAC
CGCCCACCTCCTCTGGTCCACAATGGCAGGAGATGGACAGGAGCTTTGGCCACACACAGAGGCATGTGATTCAGTCTGGGGTTCAGAATGGTTTTCTGAAGAAGGTGATATGCGAGTGAG
CCAAATCTTAAAGGATGAGTGGGCATTTCTCAAATGGTCCACAGGGTAGAGAAGAGGGCTTTGCAAAGGCCGGAAAGAGCAAGATGCGTTTGGGCTGAAACGTGCAGCCTTTGGCACACA
GAGGGGCAGAGGGAGAGCAGCCGGGATTGTGGCGCCTGTGGGGTCTGGGGGAAACAGTGTGCCCGGTGGTCACCTAGAAGCTTTGTCACAATAGTGTGGCGCCAGGATGGCTGCAGGCTT
GATGTTCACCGATGCAGGCAGCAAAGGTGCTTAGCATGGAAGAACTCCTCAAAGGGACTCGAGCTCCTCACTGCTGGGTGTGCTGAGGTCAGAAGGAAGATGCCCCCCCGAGAGAGGGAT
CAGCGGGACCCAGGCAGAGCCCTGGCTTGGACAGGCCCTGTCCTGAGACAGAGACGAGGGCTTTGCGACAATGAACAGGAATCGGGGTCCTGGGTATATGATGATGGTTAAGAACAACA
GCTTTGCAACTGGACAGCCCTGGATTTGAAACCTGGAATCACCACCTATCAGAGGGCAGTTTGTTTCAATTCTTTGAGCTGTAAGAGCTTGTTTGTTAAGCCCACAGCAGGAATGCATAA
CTGGCCATGTGTGAGGATCTGATCTTAGGACTGCGGAGCCCCGGCTGGAAGAGGCAGCCTGTGCTGAACGCAGAGATGGGTCAGGGCCACACAGGAGGCTGCTGTCCACCCCTCCTGCCC
TGAAGCCTGGCGGGGGGTCTGTGCCTTCCCGTGGAAGTGGGCTCAGAGGTGGAGATGGCATGTGTCCTGTCCTCAGAGCTCAGTGGGAGACATAGATGTAGACAAAGGATGCACAGGAGA
TGATGCCTGGGAGTATGGGGACACTGAGATGGCTGGTGGGATGCACGGCTGGAGGGACCAGGCCTGGATGGAGGAGGGGATGGGGCTGGAGGGCCTGGGTGGAGGAGGGCAGGGGGTAGG
TGCACTGTTGAGTCAGAAGCCGGAGCAGAGATGGCTTGGACTACACGGGGAGGAGCAGGGTCTTCAGAAGGACAGGGTGGGGTCTAACGTGCAGGTACATGTTGAGGGTGGGGATCCCGG
GGGACAGTGGCGTTGGAGATGTGGGGGCGGGGCTGCTCTCTGAATGGGTTTCATACTTCTGATTTTATTAAAGAAATATAAGACATTGAATTC-----(13.5 kbp)-----AGCCAAA
TACAACTGACTCTCTCTCTCTCAGGAGAGACTTAACTTTATTTTAACTTACTTTATTTTAATTCAAATTCATTTATTAATGGTTCCTTTGACAACCCGCTCACTTCTGAAATCAGAA
AACCAGCAACTTCCTCAGTTCCCACAAGTTCGCCTGGACAATCACACGGGCAGGAACAGGCGCGACACCCTGGTGGCTTCATGGTGCTTTGATTAAAGGATGTTTTATGCCACAGGATTT
CAAATGATCTCTTTGTTTGGAGTACCAGAAGTTTCTGTCCTCCAGAGAACTGCCCCTGGTACATTCTAGAAGGGTGGGAGACCCTCCTTCGGAAAACAGGGGAGTGGGAAGGAACCCTGC
AGGCAGCCACCTCCAGACAGCATGTGCTGAGCTCGGGAGAGGGTGCCTGTGGGAACCAAGCCTCTGGAATGAGCTCCCTTGAAGCCAGTCATGACCTGGGCCTCCAGGAGGCTTTCGGGT
CTTCTGGGGAGCTGTGCCCCTTGGCCTCCGGAGAGGCCCCACTTATTCTCCCTGAGAATGGTGTCTTATATCTGGAACTCACTGCTTCTGTGTTCTTCTCCCCCATCTCAAACACATCCT
                                                600                                                 650
TGCATTTGTCTCCACAG AGACCACCCCAGATGGGGCGCCTCCAACACGGCCCTGGCACGATGGCTCCCTCCAGTCTATGAGGACGGCTTCAGTCAGCCCCGAGGCTGGAACCCCGGCTTC  EXON 6
                  D  H  P  R  W  G  A  S  N  T  A  L  A  R  W  L  P  P  V  Y  E  D  G  F  S  Q  P  R  G  W  N  P  G  F
                     700
TTGTACAACGGGTTCCCACTGCCCCCG GTGGGTACTCAGAACGCTACTATCCTGGACTAAGATTGGGTCCTGTGATGCTGAGGGAGGGAAATAGCCTTTTACTGGGTTCTTGCTCGTGCT
 L  Y  N  G  F  P  L  P  P
GGGTGCTGCGTGCAATCCCTTCAGATCCTCCCAGCCTCCCTTTGATGTAGCAATCACTGTTTCTGCCCTATGGCTTAGGAGCCAAGGCTCAGGGAGATTGAAATCTCTTAGTGAGTGGAA
CCCTGCAGATTTAAATTAAGTTTTCTCAACCCTAACGTTTCTGTTCCTGAAAATCACTTCAGTGTTTTCAGATAAATTCCTTTATCCATCATGTGGGCTAGGTACATCCCAAGATGATTA
TGAGTTCAGCAATCCACTTGTTGATGTAAAGATGACTGCTTTACTCCCAGAGGTGAATTTATTTGCTGGACATTCATAGCCTAGTAGAATCTACTGGTCTATTTGGTCAAAAAATCGGTC
ATTTTGCTTGTAATTTAACTACTCATTTCAATGAATAAGCCAGCCTGTTTGCTCTTGTGCCATGTTCAGTGATGAAATTATCTAGAAATACATCAGTGGATGTTCTGTTTAGCCCAAAGA
GCGCAGGTGGTGGGACTCAGGTCTGACTCAGACCAGCTGTGGACAGATACACCAGCACAGAGTTCCCCTACGAAAGCCAACAACTTTACCTAGGCAATTTTCCTAGGCGGGCTTTTAGTA
AGAGGACCTGGTGTTCCCAGAAGCCGTGAGAAGCCCTGGGTAGAGCTGGAGGCTGTGTCACTGAAGACGGCTCGGTGCTCCCAGCAGCAGGGACAAGGCAATTGGGCTCAGCTCTGCTCC
CACTGGCTTCCTGCAAGTCGCTTGCCTTCTTGGCCTCAGTTGGACAATCCCTAAAGTGGGATAACACTGTCTTACCCACCTTATGCTCAACAGTGGACCAGTTCAACCCCCGAGGCTTTT
TAAGCTCTCAGTATTATGAGATTGTGATTCTGTAAACCCATATTTGCTCAGAATTATTTCACTATTGAGAAAAATCAGAGTTCAAGTGTCTTTAAATGAAATTGTGTTGGTTTTCTACCC
AGCTGTGACAAATTGCTTCAAATTTAGTGGCTGAAAACAACATAAATGTGATCTTCCAATTTCACAATCAGTGTCTGATGTCTCACTTGGCTGAAATCAAGGTGTTGCAGAGCTGGTTCT
TTCTGGAAGCTCCAGGGGAGATCCCTTCCCTGGCAATTCCAGCTTCTGGAGGTGCCTGCATCCCCTGGCTCATGGCCCCTCCTTCCAAGCCAGCAATGTCTGTCTCTGTGCCTTTGTT
TGGAGTCACATCTCCACCTCCCGTCAGAGAGGGCCTCCAACTTTAAGGACCCATGTGACTGGATTGGACACTTGTGGAGACCCAGGACAACCTTCCCATCTCAAGGCCCTTAATGTAATC
GCATCAGCAGGCTCCAAGGATTAGCGAATGGATGTTGTGCGAGAGGCATTTTTCTGCCTCCACAGATGACCCACTGACCCTCAAAGATTCATACCCAGTCCACATGCCTAACATGCTCAC
CCCATTCCAAGGTTCCCCCAAACCCCATCCCATGACGGCACCAACTCTAGGCCCAAAATCCAATCCAAATCTCATCACATAAATCATCCACATCAGGTAGGGCAGGCTCAGGGTAAGATC
CATCCAGGACAAAATTCCTCTCCATCTGGACCTAGCCTGTGAGACAGGGAAGCTTGTCGTCTGCTCTCCAGATGCCCTGGCGGAACGGTGTTAAACATAACAGCAACAGATGTCTGCGTG
GAGAATGGAAAGGAAGGGGAGTCACTCCTCCCAACCCACGTGGAAATCCAGCCAAGCAACTCCCTTAGGCTTCAAGGCCAAGGAGTACTTGCTTGTGGCTGGAGGCTCTGCCTCTGAGTC
ATGAACTCAACTGTCCTAAAGCCTGGCTGTTTTGTCAATGACTCTTCACTTACCCCACAACAACACTGAAAGCTCAACCTTCCATATTAACTTACAGATGAGGAAGTGAGGGCTGAGGGA
AGTTAATGCCAGGCCCTCCAGCCACCACAAATCAAGGCAGGGATTCAAGTCCAGGGGCATCTGTCTTCAAAGCTTATGCTCCTCCTGCCCTAACTCCAGGGCACAGATCTCCTAGGGGCA
CCTGGAGCTCTGTGAACAAGAACCACACCAGGAAGTGCATGATCCCAAAGGTCATCTTTCTGCTACCACGGGGTCCTCCTATGTCCTGACCAATGGTCTCTTCCTACCCAG GTCCGGGAG  EXON 7
                                                                                                                    V  R  E
                                                750                                                 800
GTGACAAGACATGTCATTCAAGTTTCAAATGAGGTTGTCACAGATGATGACCGCTATTCTGACCTCCTGATGGCATGGGGACAATACATCGACCACGACATCGCGTTCACACCACAGAGC
 V  T  R  H  V  I  Q  V  S  N  E  V  V  T  D  D  D  R  Y  S  D  L  L  M  A  W  G  Q  Y  I  D  H  D  I  A  F  T  P  Q  S
                850                                                 900
ACCAGCAAAGCTGCCTTCGGGGGAGGGGCTGACTGCCAGATGACTTGTGAGAACCAAAACCCATGTTTTCCCATACAA GTAAGTTTTAGAAAACGTTTCTATGTTTGTTATGAAACTAAG
 T  S  K  A  A  F  G  G  G  A  D  C  Q  M  T  C  E  N  Q  N  P  C  F  P  I  Q
TGCTATTTAAAACGTCAAACAGAATGGTATAAAACAAAATGTGAAAGTCTGTTTCTTTGTCCTATTCCCAGGGGTAGCAGTTGTGAATATCTGGTGCATCTCACTGGAAATTCCCTGGGT
ACACGTATATATAGCATGTATGATAGTGTGGGTACACATATATATAGCATGTATGATACTGTGTGTACACATATATATAGTATGTATGATAGTGTGTGGCAGATGTGTACATAGCATGT
ATGATACTGTGGGTACACATATATATAGCATGTATGATAGTGTGTGGGCAGATGTGTACATAGCATGTATGATACTGTGGGTACACATATATATAGCATGTATGATAGTGTGTGGGCAGA
TGTGTACATAGCATGTATGATACTGTGGGTACACATATATATAGCATGTATGATAGTGTGTGGGCAGATGTGTACATAGCATGTATGA---------(?)--------CCGGGGATCTCC
ATGTTACAGATGAGAAGATGGAAGCTGGCCACAGCCAGAGTTGGAATCCCACCCAAGGTCTCCTCACTCCCCTCAAGCTGCCTTTACACTAAACAGTGATGCCCACAAAACTCGGAGACT
GTTTACTCATTGTGCTCCTAAACCAAGTCCATGGCACTGGAAAGGAAGAAGTCAGCTCTTCCGAAAGTCAGAAAATAAACCAGACCGGGGATTCCATGAGCAGCTAGAGCAGGTCTGAGC
AGGAGCGAGACTGGCTCAGGCCAAGAGCGTCTGAGCTTTGGGGAGATGTGGACTCCTAGGATGTGGGGTCCATCCTCTGGGTTCAACCTCTTCCAGGCAAGAGACCACCAGTGAGCAGGC
TGAGGGGAGGTGCGGGCTGGCTGGACCGACCCCTGCAGCCTCTTCCAGGCAAGAGACCACCGGTGAGCAGGCCAGGGGGGAGGTGGGGGGCTGGCTGGACCAACCCCTGCAGCCTCTTCC
AGGCAAGAGACCACCGGTAAGCAGGCCGGGGGGGGAGGTGAGGGGACTGGCTGGACTGACCCCTACAGAGGACTGGAGGGGCAGAGAAACGTGCGGCGCTGCGGGGTCGTCGCCGGCCTC
                                                                                                                    950
GAACTTCCAGAGTCTTACAAAGGGTGCACGGGGGCCCTGGGTGACCTTGAACTCCCCTTTGCCTGCAG CTCCCGGAGGAGGCCCGGCCGGCCGCGGGCACCGCCTGTCTGCCCTTCTACC  EXON 8
                                                                                                L  P  E  E  A  R  P  A  A  G  T  A  C  L  P  F  Y
                                                1000                                                1050
GCTCTTCGGCCGCCTGCGGCACCGGGGACCAAGGCGCGCTCTTTGGGAACCTGTCCACGGCCAACCCGCGGCAGCAGATGAACGGGTTGACCTCGTTCCTGGACGCGTCCACCGTGTATG
R  S  S  A  A  C  G  T  G  D  Q  G  A  L  F  G  N  L  S  T  A  N  P  R  Q  Q  M  N  G  L  T  S  F  L  D  A  S  T  V  Y
            1100                                                1150                                                1200
GCAGCTCCCCGGCCCTAGAGAGGCAGCTGCGGAACTGGACCAGTGCCGAAGGGCTGCTCCGCGTCCACGCGCGCCTCCGGGACTCCGGCCGCGCCTACCTGCCCTTCGTGCCGCCACGCG
G  S  S  P  A  L  E  R  Q  L  R  N  W  T  S  A  E  G  L  L  R  V  H  A  R  L  R  D  S  G  R  A  Y  L  P  F  V  P  P  R
                                                1250                                                1300
CGCCTGCGGCCTGTGCGCCCGAGCCCGGCATCCCCGGAGAGACCCGCGGGCCCTGCTTCCTGGCCGGAGACGGCCGCGCCAGCGAGGTCCCCTCCCTGACGGCACTGCACACGCTGTGGC
A  P  A  A  C  A  P  E  P  G  I  P  G  E  T  R  G  P  C  F  L  A  G  D  G  R  A  S  E  V  P  S  L  T  A  L  H  T  L  W
                1350                                                1400
TGCGCGAGCACAACCGCCTGGCCGCGGCGCTCAAGGCCCTCAATGCGCACTGGAGCGCGGACGCCGTGTACCAGGAGGCGCGCAAGGTCGTGGGCGCTCTGCACCAG GTGCGCGGGGTGG
L  R  E  H  N  R  L  A  A  A  L  K  A  L  N  A  H  W  S  A  D  A  V  Y  Q  E  A  R  K  V  V  G  A  L  H  Q
```



```
AGTGCAGGGAGGTTGCTGCCTCCAGCCTGGCCCTCGGCGGAGGCTGCACCCAGGCCAGCTCTGTTCCCTCCAGACAGAGCAGCTCCTGGAAGAGCTCGTGTTCCCCACACTGACAGCACC
TGCTCCATTTAGAGAATGCGCACTGGGCATGTAGAGAAGACCTGTGACACGCTACAGCTTCTTTTCTAAAACTGGCCAGAAAAGGAACATTATTTGAACAAAATGTACAGCAGGTTATCT
GGAAACATACGCCTAGGGGATGAGATACTAGGTTGTAGGATTCTTGTGCAGGGGGAGCAGGAGCAGGAGGGGACCCCACATCCTCACTGGCGGTGGCAGGTGGCCCCAGTCCTTGCTCTC
AGAGGACCTGCCGGCCAGCTGGGAGCATCTGCCTGGGCGAAGCAGACCCTGCCCCACCCTCACCCTGGCCCCTTCCTGCCCCCGTGTCCTGCACTTGGCCAAATTCCACATTCCCAGGCA
GACCCTGATTTCCTCCAGGAAGTCTCCTGTGACTCGGGCCTGAGTGAGCCCCAGTGTGTGGACTCCCTGCAGGCTCAGAGCCCGTGGCCAGCGCACATCTGTCTGCTCCTCATCACCTTT
TCGGATGTGCCGGGAGCAGGGGGCGGCCGGTTCCCCCTAGACCAGGTGGGATGGCAGGCAAAGGGCTGACCTCCCCAGAGAGAAGCACCTCCCAGAACGGGGGTCGCTCGCGGGAGATGG
                                                                    2500                                                2550
GGGTGCAGCCGCTTCCTCTCACGTGTGTGGCCTTGTGTGTCTGGCAG ATGTGAACGAGTGTGCAGACGGTGCCCACCCCCCCTGCCACGCCTCTGCGAGGTGCAGAAACACCAAAGGCGG  EXON 14
                                                  D  V  N  E  C  A  D  G  A  H  P  P  C  H  A  S  A  R  C  R  N  T  K  G  G
                                                                2600
CTTCCAGTGTCTCTGCGCGGACCCCTACGAGTTAGGAGACGATGGGAGAACCTGCGTAG GTGAGGCTGTTCCCCTCTCTCTCTGTCCGTCCATTTGCGAGTTTTTGTAAAATGAAGAAAA
  F  Q  C  L  C  A  D  P  Y  E  L  G  D  D  G  R  T  C  V
GTGTGTTCAAGTTAGGCAACTGTCAATCACCATCTTGATTGGGAAGGGCGGAGATGAATAAGACACCAAGCCCACGGTGCCCGAGGGGCGTCTGCGCTGTGCTGGGGGCAGAGGATGCAC
TTCGTGTGCCTGCTGTCTGCAGAGACCCTCGAAGCACACTGAGGCTGGGGACGGCAGCCAGAATGCATTGCCAGACAAGGCCCACAGGTGTATACCCAGCCTAGATCACACCCCAGACAA
AAAACCTATGAAACAGGCAGGATTCATTACAGACCCATCCATCACCTCTCCTCCTCC-----(11.0 kbp)-----GAATTCCCACCCATGACCCATATCCCCTTCCCCACACCCCCAG
CCCCAGGCACACAGGCGGACCCACCTCTAGCAGCCATGGGGTGAGAGCGAGTAGCAGCAGATGGAAACAAGAACCCAGCCTCTTCTGGATGCAGGAGGAGCGGCATCCCAGGGCCAGCAA
GCAGCTCTGCTCCCCATGGGCTTTTAGCAAGAGGCTGGTGTGCTCAGGTTGCCCTGCAGACTCCCCTGCGTCTGTGAAAGGTGGCAGGAAGGAGATCCTGGAGGCGGGAGGAGTGGTTAGA
GGAGCCCAGGAGGAAAAGTCATGAGGTGCAAAGGCTAAGAGGACAGGATCGTTTTGGAGAAACAGTCAGGATGCGCATTTAGGGTTTATTCATTCATTCTTGTATCCAACAAATGCTTGC
TGAGTGTCCAAGACCCATGCGAGGAAGACAGCCACGGACACCACCTGTGTCTGGGCTGGGAAAGACCCATGATAGGAAGACAGCCATGGACACGACCTGTGTCCGGGCTGCAAATGAGAC
ACACCTGCAAATCATGCCCAGGTGCTAAGTGGCACAGGGAACCGCAGTCACCCACAGACGCATTCATCACCGACTTTAAAAAATCCAGTGTGAGAAATAGAACCAAACCAAGCACCCATT
CCCCCAGCCGCTGCCTCTGTCTGGTACCTGAATGTGCTCTCTTAGCCCCAGAGCCTATCTTCAACCACACGTGATGATGTTTATCCAGGGCGACCCATGTTTGTTTAAAAAAAAACGCT
GATTGGTAGTATTTAAATTCTTATTTAAAAGTCAGTTTTCGTTTCATCTTGGATTTCTCCTCGGCCACCTCGAATGGGTATCCAAACAGAAAAGTCGGTAATAAAACTGTGGTCAAAGGA
GCTCTGAGTTCAGAGCCGCAGTGGCCCAGGCTGGCAGCTCCTTGCAGAAGCACAGAAGCCAAGGCTGGAGCCCAGGAGCCGTTACTCAGTCCCCGGCTTTTTCAGATGAAGACACAGGTC
CAGGTGGGAGGCACTTCCTCAGGCCCCTCACCAGATGCTCTGGGGCCGGCTCCTGGGCTGCCTCTCTCCTCTGTCCCATCAGGCCTGCACTTCTCTGTCTGATTCCCACCAACCTTGTTC
GGTTTCTTTCTGGAATTGGCTGATGAGACCCAGGATGAGCTCCAAAATCTGCTGCCATCCCGTGCCCCCGTCCCCCCAGGGCTCCAGCTCTGCAAGGTGCTGTCCACCTTCCTTCAACCA
CGGTTTCACCTCCAGGAGGCTTCTGGGAAAGATAAAATGCAGCCAAGGGCAAAACCTCACAGCGTTATACCGCGTCCCTTACAGCCCACCTTGCGAACCTTTCCCGCCCCGGTCCGAGGC
CTCCAGGGCTCCCTGGGTCCTGTGGTCCACTCCGAGGGCACTGAGAGCCCTCCCTGGTTTGGGCAGCAAGGTGTGCTTTCTGTCCTGCAGGACGCGTCCATGCCGAACCCACAGTGAGGT
TCCATAGAGAGAGCTAACTTCTCTGCCCTGTGTCCCCTGCTGGCTCTCCCCACACACGTCACATTCGGATCTGGACTTAGTGGTCTAAATGGCTTTTCTGGGCAGATAAATGTCCCTCCC
TCCCTGCGTCATGCTGTGGGGTGAAGGCCACTGGCCAGGACAGGGTATGGCCCAGACTCAGGCAGGACAACCTGGCTTGCCCAGGCCCTGGAAGGTTCTTCTAACCAGGCCTCTTTCTGT
                                                                2650                                                2700
GCCCCAG ACTCCGGGAGGCTCCCTCGGGTGACTTGGATCTCCATGTCGCTGGCTGCTCTGCTGATCGGAGGCTTCGCAGGTCTCACCTCGACGGTGATTTGCAGGTG GTAAGTCCTTCAC EXON 15
        D  S  G  R  L  P  R  A  T  W  I  S  M  S  L  A  A  L  L  I  G  G  F  A  G  L  T  S  T  V  I  C  R  W
TTTTTGACTGTTACTTAGACACAAAGCAATCTCCTTTCCTCTGGACATGGCTGCAATGAAGCTGCTTTATTTTCCCGAAGCTAACACAGGTTACTGGTATCTCAAGGCATTGATCTCAGA
GCTATTTTGTACAACGTAATAGACTCTCCGGATCCGTT---------(?)---------TTTGCTGGCGGTGCTGGAATGGCTATGCGGAGGGCGCTGTGGTGCGGGGACTGCGCTCCCCA
GGAGGACCTCGGGAGTTGTTTGTTTGTAGGTATTACAGTATTGTCTAACGTTGCCGCGTTGACACAGGGTCACGGGAAGCCACCTCCAGGGCCTGTGGTCCTCTTCAGGCAGGCGGGGCG
GGGAATGCTCAGGACCGGTGATGATGGTGCCATGCAGAGTTCAGGGATGAATGACCACCAAGGCAAGGGAAGGAGTCCAGCAGCTCAGTATTTGGGGAATGGGGGCGGGGGCAGGTAAAA
ACGCAGAATCAAGGGAAACAAACAGCTGGAAAATATAGTAGCTGTTAAGTTCTGATGGAAAAGCTGTTTTGTACTTACAGTTAAATCCCACAGTATTTGTGATGTATTTCCGGGTAACTT
GCTCTTCCCTCCTCAAAAGGTCCCAGCCTCACCAGAGCTTCCCTTCCTGCACCGGCCCCAGCTCTCCACGCCTCCCAGAGACCAGGGGGCCTCCACCCTTGCTCCTCTGGGACGGCGCTT
CCTCCCCCTCATCTCCAGGTCGGGGCTGGGGTCTTCAGGCTCAGGTTGAAGATGCTGTTGGGCCCTTTGCCTGTACAGGGATGGCGCCCGACCCAGGAAAAGCCTAACGTTGTCTGTCAT
TTGGATGCGTCTCTTGGCAGTGACTGGGGGGAAATCCTCACTGCAGCCCGCGTCCCTTGCATGATTGTTGTAAGAAAGCCCCAGCTCAGCTGCGTGGAGACAGGGTCCTCTTGGCTGCAA
ATTCTAAAATCATTTTTCCTATGAAGAGAGCAGTGCTAATTTTTCCAAAATATATCAGATTATGATCGACTTGACTGAAGTGTGAAATGAAAGTGGGTTGGAGTGTTCCTGCCAAAGACA
AGCACGGCTGCCTTGCCGTCGCTCGTGCCGTGCTCTCTACCCTCCACAGTCACGGTGCCGGACCCTCTCCCGATAACTGGACACGTGTCTCCCACAG GACACGCACTGGCACTAAATCCA  EXON 16
                                                                                                    T  R  T  G  T  K  S
                2750                                                2800
CACTGCCCATCTCGGAGACAGGCGGAGGAACTCCCGAGCTGAGATGCGGAAAGCACCAGGCCGTAGGGACCTCACCGCAGCGGGCCGCAGCTCAGGACTCGGAGCAG GTGGGCCACACCA
T  L  P  I  S  E  T  G  G  G  T  P  E  L  R  C  G  K  H  Q  A  V  G  T  S  P  Q  R  A  A  A  Q  D  S  E  Q
TGCCGCATGTTTCCAGCTGCCACCGCAGTGGTTGGACAGGATCTGGGTGTCGGAGCAGCTCTGCTGGGGCTCCCTGCATATTTCTGTTTACTCCGTGTTTCCTAGTCCGTTCTGCACCTT
CCTCCGGGAGGTTTTATTTACAAGCTAATGACAGTGAGCCAGAATGTGAGCCTGCTTAGTGAGAGGAATGAAGATGCCTTCCATTTGTACAAACCGGTTCCAAAACTAAGGGCTCACTTC
CAGGCGTTTGCTTCCACTTAATCTGAGGATCGGCTGGAAGCACAGGAGAGGAAGGAGAAGCTGAAGCAAAACCCTGACTTTTAAATTCTGATTTTTAAGTGGAATAGTTATAATCAGTGT
GGAAAAATGTGTATGTTTATGTTTACCCCCTTACCTTGTATGCAGAACCCCAAGGGAGCCATATCAAAAACTCACTTGTGTAAGTCTGACTTTGAACTGAGAGAAGCAGAGAGCAGAAC
AGTGGCCTCCAGAGGGGAGTGGGGAGGTGCTGGTCACAGGCACAAAGTCTCAGGCAGGAGGAGGGCGCCTCAGGTCTGCTGCACGGCATGGGGCGCACGCTTCCTCATAATGATCGCATA
TTTCAAAATTGCTAAAACAATAGGTTTTTTTTTTTTTGAGATAGAGTCTTGCTCTGTCGCCCAGACTGGAGTGCAGTGGCACAAACATAGCTCACTGCAGCTTCGACCCTCCTGCTCAAG
CAATCCTCTTGCCTCAGCCTCTTGCGTGGCTGGGTCTACAGGCGCTCACCACCATGCCTGGCTAATTTTTAAATTTTTTACAGAGATGCAGCCTCGCTATGCTGCCCAGGCTGG------
                                                                2850                                                2900
(1.9 kbp)-----GAATTCAGACGTTATTAATGTTTGTTCTGCATTTTTGCAG GAGAGTGCTGGGATGGAAGGCCGGGATACTCACAGGCTGCCGAGAGCCCTCTGAGGGCAAAGTGGC  EXON 17
                                                                        E  S  A  G  M  E  G  R  D  T  H  R  L  P  R  A  L  *
                                                2950                                                3000
AGGACACTGCAGAACAGCTTCATGTTCCCAAAATCACCGTACGACTCTTTTCCAAACACAGGCAAATCGGAAATCAGCAGGACGACTGTTTTCCCAACACGGGTAAATCTAGTACCATGT
                3050
CGTAGTTACTCTCAGGCATGGATGAATAAATGTTATAGCTGC ATTTGTCTGGCCTTTTCTTGTAAACATTGCCTGATTTGTTCCTTCTGGGGCTTTGCATTAAAATGTATTTACAGATTA

CACATCTTTATTTTGTGAACCCTGGAAACACCACTCTTGCAATCCTCCTGTCTCCACCTTCTGGCATCTCTGATGCGTGCTCGTCTGCACTCTGCCCCGGCGGTCCCTCCAGCACTGGTT
TTTCCACACCCCCTGCCCATCACCAGGTGTCCACAGGGCCTACATCTGGCTTCCCTCTTCTCCTGGGAAGAGCACTCCTGGCTTCCTGCAGGGCCGGTGGGAGGAGGAAAGTGATTCTGA
GGGAGAGGCTGGAGCCTTAAGGACCAACAAGGCAAAGTGACTTGTCTCATCCTCCAGAGATTCACCAAACACATGAGCCTCAGACCCCAGGCTTCTGCCTCCAGCAGCCGCCCTGCCGCAC
ACTGCTCTTACTCCTCCTTATACCCTCACTCACGGGGAACACAGCCCAGTGATCCCGGAGGAAACTCACTCCCTCCCTGACTCAACAAGGCAGTCTCGGGGGCACCGTTAGCCACGCGAC
CCTGTAAAGCTGCCGTCCTCATTTCACATGTGAAGCAGCTGAATTC
```

FIGURE 3: Sequence of the human TPO gene. The sequence of the structural gene is displayed beginning at +1. Upstream sequence begins with a −1, immediately before the major start site. Only the exon sequences are continuously numbered, and the location of each exon is indicated at the right. Gaps (---) represent introns whose sequences were not completely determined. Estimations of unsequenced intron DNA were based on sizes calculated from restriction maps in Figure 1 and the size of the intron sequences that had been determined. These are shown in parentheses except introns 7, 10, and 15 in which the sizes are not known (marked by ?). The TATAA box and the poly(A) addition signal are *underlined*. Nucleotide sequences coding for the predicted proximal and distal histidine residues (Kimura & Ikeda-Saito, 1988) are *bold-overlined* and *overlined*, respectively. The termination codon is marked by an *asterisk*.

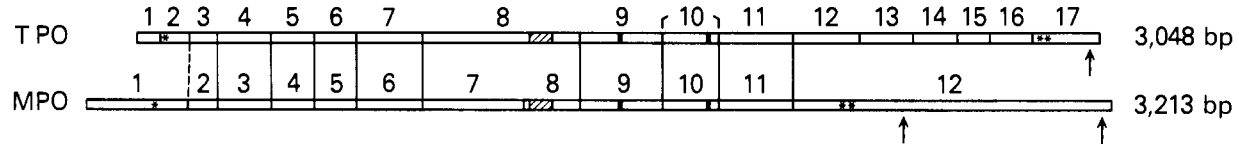


FIGURE 4: Schematic comparison of the cDNA structures between human TPO and MPO. TPO and MPO cDNAs are divided into 17 and 12 exons, respectively. The positions of homologous exon–intron junctions between TPO and MPO are shown by a *solid line* extending between the genes. The second exon–intron junction in TPO and the first in MPO are not precisely aligned and therefore are denoted by a *dashed line*. The initiation codon is marked by an *asterisk* and the stop codon by *two asterisks*. The poly(A) addition signal is shown by an *arrow*, and the alternatively spliced sequence in TPO (Kimura et al., 1987) is marked by a *bracket*. The *hatched box* and the *black boxes* show the conserved putative proximal histidine-containing sequence and the possible distal histidine-containing sequences, respectively (Kimura & Ikeda-Saito, 1988).

The relation of the exon–intron junctions of both peroxidases is summarized in Figure 4. Alignment of the TPO and MPO cDNA nucleotide and amino acid sequences (Libert et al., 1987b; Kimura & Ikeda-Saito, 1988; Morishita et al., 1987b) revealed that the sequences, corresponding to exons 3 through 11 in the TPO genes, displayed significant similarities to those

Table I: Comparison of Exon–Intron Junctions in the Human TPO Gene with Those of the Human MPO Gene[a]

	Exon		Intron		Exon
TPO	1	---TCAG/GTTGT---	1	---TTTAG/AATG---	2
		A(52)		Ia(52)	
MPO	1	---CCA—/—C---	1	---G C /CTGT---	2
TPO	2	---TGGG/GTAAG---	2	---CGTAG/GAAA---	3
		G(32)		Iy(32)	
		Se(83)		r(83)	
MPO	2	--- A—/—G ---	2	---AGC /C T ---	3
TPO	3	---AGAG/GTGAG---	3	---CATAG/AAAC---	4
		Ar(60)		g(60)	
		A(142)		sp(142)	
MPO	3	--- TG—/—C ---	3	---GG —/— T---	4
TPO	4	---ACGG/GTAAT---	4	---CACAG/ATGC---	5
		A(117)		sp(117)	
		Ar(183)		g(183)	
MPO	4	--- —/—GCG---	4	---TG —/—CG ---	5
TPO	5	---ACAG/GTATT---	5	---CACAG/AGAC---	6
		Ar(161)		g(161)	
		Leu(226)		Ala(227)	
MPO	5	---T—T—/—A ---	5	--- G —/—CT ---	6
TPO	6	---CCCG/GTGGG---	6	---CCCAG/GTCC---	7
		Pro(204)		Val(205)	
		Lys(295)		Ile(296)	
MPO	6	---CA G/—GGC---	6	---CT —/—A ---	7
TPO	7	---ACAA/GTAAG---	7	---TGCAG/CTCC---	8
		Gln(273)		Leu(274)	
		Gln(455)		Ile(456)	
MPO	8	--- —/—AG ---	8	---TTT —/— ---	9
TPO	8	---CCAG/GTGCG---	8	---CACAG/ATCA---	9
		Gln(446)		Ile(447)	
		G(541)		Iy(541)	
MPO	9	--- A—/—A ---	9	---C —/— ---	10
TPO	9	---GGAG/GTGAG---	9	---TCCAG/GTGG---	10
		G(533)		Iy(533)	
		G(598)		Iy(598)	
MPO	10	--- —/—GA ---	10	--- —/—A ---	11
TPO	10	---CCAG/GTCTG---	10	---TGCAG/GTTA---	11
		G(590)		Iy(590)	
		Ar(677)		g(677)	
MPO	11	--- TC—/—GA ---	11	--- C —/— ---	12
TPO	11	---ACTG/GTACG---	11	---TGGAG/GTTT---	12
		Tr(669)		p(669)	
TPO	12	---CAAG/GTGAA---	12	---CATAG/ACGA---	13
		A(739)		sp(739)	
TPO	13	---AAAG/GTCAG---	13	---GGCAG/ATGT---	14
		A(798)		sp(798)	
TPO	14	---GTAG/GTGAG---	14	---CCCAG/ACTC---	15
		A(840)		sp(840)	
TPO	15	---GGTG/GTAAG---	15	---CACAG/GACA---	16
		Tr(873)		p(873)	
TPO	16	---GCAG/GTGGG---	16	---TGCAG/GAGA---	17
		Gln(916)		Glu(917)	

[a] The nucleotide and amino acid sequences at the exon–intron junctions of the TPO gene are compared with those of the MPO gene. In the case of MPO, only nucleotides that differ from those of TPO are displayed.

of exons 2 through 11 in the MPO gene. However, the junction between the seventh and eighth exons of the MPO gene does not have any counterpart in the TPO gene. The intron junction separations of the amino acid codons are very well conserved between both enzymes (Table I).

The intron loss hypothesis was proposed to explain the role of introns during evolution (Cornish-Bowden, 1984; Davie et al., 1986; Doolittle et al., 1986; Serrapathy, 1986; Gilbert et al., 1986; Traut, 1988). In this scenario, introns were originally present in the first genes and were lost during the evolution of prokaryotes and other organisms in the interest of a more effective and rapid means to replicate their genomes. Less DNA content due to lack of introns would allow rapid DNA replication and quick cellular division. A conversion hypothesis is that ancestral genes resembled prokaryotic genes and that addition of the introns took place either concomitant with the emergence of the eukaryotic cell or gradually in the stream of evolution (Rogers, 1985; Traut, 1988). Assuming that the intron loss occurred during the evolution of peroxidase, the ancestral peroxidase gene probably contained at least 12 exons (Table I, Figure 4).

While 9 exon–intron junctions have been conserved, the intron corresponding to intron 1 in TPO was probably lost during evolution of the MPO gene, although the possibility of intron insertion in TPO cannot be ruled out. The positions of second TPO and first MPO introns appear not to be conserved, probably resulting from a high divergence in this portion of the nucleotide sequences between both enzymes. Alternatively, it is possible that intron insertion might have taken place independently in both peroxidases. Moreover, the seventh intron in MPO does not have a counterpart in TPO, further indicating that either intron deletion in TPO or intron insertion in MPO has occurred. It is also interesting to note that although the overall amino acid sequence similarity between TPO and MPO is 44%, the similarity is higher in the sequence coding for the mature fully processed MPO protein, which is coded by the latter half of the MPO exon 4 (Kimura & Ikeda-Saito, 1988). Furthermore, the seventh intron of MPO is located just prior to the predicted proximal histidine-containing region which has 74% similarity between two peroxidases. Whether intron deletion or intron insertion events played important roles in the evolution of TPO and MPO genes should become more evident once other peroxidase genes that share the same ancestral gene as TPO and MPO are isolated and sequenced, if indeed others exist.

Exon shuffling (Rogers, 1985; Davie et al., 1986; Doolittle et al., 1986; Gilbert et al., 1986) has also played a role in the evolution of the peroxidase gene family, for example, in the origin of exons 13, 14, 15, and 16 in TPO. In this regard, it is interesting to note that the percent nucleotide sequence similarity between the TPO and MPO coding regions drastically decreases immediately after the MPO termination codon (where the TPO exon 13 starts). The first poly(A) addition signal in the MPO sequence (Johnson et al., 1987; Morishita et al., 1987a) was found about 160 bp downstream of the stop codon; a similar distance between the poly(A) addition signal and protein stop codon is also found in TPO. As Libert et al. (1987b) have proposed, and the results in Figure 4 and Table I show, four exons coding for different protein domains, including the C4b-β2 glycoprotein, the EGF-LDL receptor, and the transmembrane domain, were inserted into the ancestral TPO gene during evolution, resulting in the present TPO gene. The addition of other exons, possibly derived from other genes, into the TPO gene and the intron loss and/or insertion found in both TPO and MPO genes all took place after two peroxidase genes separated and started their own evolutionary pathways.

Finally, of more interest, is that in spite of both cDNAs having similar length, the size of MPO gene is only 10 kbp whereas that of TPO is probably at least 10-fold longer even if the 3′ portion of the TPO gene which was added later only to the TPO gene is not considered. In this regard, it is noteworthy that the human thyroglobulin gene, a highly specialized gene, is expressed only in thyroid tissue and is over 300 kbp in length (Baas et al., 1986). The reason for the large size of genes specifically expressed in thyroid is presently unclear.

ACKNOWLEDGMENTS

We thank Dr. Frank J. Gonzalez for his help and useful

discussions throughout the course of these experiments and for the preparation of the manuscript and Dr. Harry V. Gelboin for support and encouragement.

References

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Smith, J. A., Seidman, J. G., & Struhl, K. S., Eds. (1987) *Current Protocols in Molecular Biology*, pp 4.6 and 4.8, Wiley Interscience, New York.

Baas, F., van Ommen, G.-J. B., Bikker, H., Arnberg, A. C., & de Vijlder, J. J. M. (1986) *Nucleic Acids Res. 14*, 5171–5186.

Cornish-Bowden, A. (1984) *Nature 310*, 724.

Czarnocka, B., Ruf, J., Ferrand, M., Carayon, P., & Lissitzky, S. (1985) *FEBS Lett. 190*, 147–152.

Davie, E. W., Ichinose, A., & Leytus, S. P. (1986) *Cold Spring Harbor Symp. Quant. Biol. 51*, 509–514.

DeGroot, L. J., & Niepomniszcze, H. (1977) *Metabolism 26*, 665–718.

Deininger, P. L. (1983) *Anal. Biochem. 129*, 216–223.

Doolittle, R. F., Feng, D. F., Johnson, M. S., & McClure, M. A. (1986) *Cold Spring Harbor Symp. Quant. Biol. 51*, 447–455.

Gilbert, W., Marchionni, M., & McKnight, G. (1986) *Cell 46*, 151–154.

Heilig, R., Lemaire, C., & Mandel, J.-L. (1987) *Nucleic Acids Res. 15*, 9129–9142.

Johnson, K. R., Nauseef, W. M., Care, A., Weelock, M. J., Shane, S., Hudson, S., Koeffler, H. P., Selsted, M., Miller, C., & Rovera, G. (1987) *Nucleic Acids Res. 15*, 2013–2028.

Kimura, S., & Ikeda-Saito, M. (1988) *Proteins: Struct., Funct., Genet. 3*, 113–120.

Kimura, S., Kotani, T., McBride, O. W., Umeki, K., Hirai, K., Nakayama, T., & Ohtaki, S. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 5555–5559

Kotani, T., Umeki, K., Matsunaga, S., Kato, E., & Ohtaki, S. (1986) *J. Clin. Endocrinol. Metab. 62*, 928–933.

Leach, D. R. F., & Stahl, F. W. (1983) *Nature 305*, 448–451.

Liang, J. C., Chang, K. S., Schroeder, W., Siciliano, M., Trujillo, J., & Stass, S. (1987) *Am. J. Hum. Genet. 41*, A226.

Libert, F., Ruel, J., Ludgate, M., Swillens, S., Alexander, N., Vassart, G., & Dinsart, C. (1987a) *Nucleic Acids Res. 15*, 6735.

Libert, F., Ruel, J., Ludgate, M., Swillens, S., Alexander, N., Vassart, G., & Dinsart, C. (1987b) *EMBO J. 6*, 4193–4196.

Magnusson, R. P., Chazenbalk, G. D., Gestautas, J., Seto, P., Filetti, S., DeGroot, L. J., & Rapoport, B. (1987) *Mol. Endocrinol. 1*, 856–861.

Maniatis, T. (1978) *Cell 15*, 1157–1174.

Morishita, K., Kubota, N., Asano, S., Kaziro, Y., & Nagata, S. (1987a) *J. Biol. Chem. 262*, 3844–3851.

Morishita, K., Tsuchiya, M., Asano, S., Kaziro, Y., & Nagata, S. (1987b) *J. Biol. Chem. 262*, 15208–15213.

Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459–472.

Nunez, J. (1980) *The Thyroid Gland* (De Visscher, M., Ed.) pp 39–59, Raven, New York.

Ohtaki, S., Kotani, T., & Nakamura, Y. (1986) *J. Clin. Endocrinol. Metab. 63*, 570–576.

Portmann, L., Hamada, N., Heinrich, G., & DeGroot, L. J. (1985) *J. Clin. Endocrinol. Metab. 61*, 1001–1003.

Rackwitz, H.-R., Zehetner, G., Frischauf, A.-M., & Lehrach, H. (1984) *Gene 30*, 195–200.

Rogers, S. J. (1985) *Nature 315*, 458–459.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.

Serrapathy, P. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 2133–2137.

Traut, T. W. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 2944–2948.

Weil, S. C., Rosner, G. L., Reid, M. S., Chisholm, R. L., Farber, N. M., Spitznagel, J. K., & Swanson, M. S. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 2057–2061.

Wyman, A. R., Wolfe, L. B., & Botstein, D. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 2880–2884.