The extracellular hemoglobins (Hbs) of annelids and tube worms are giant multisubunit proteins of up to \( \approx 200 \) polypeptides and molecular masses to at least 3,900 kDa. They differ from all other Hbs in having both \( \beta \)-binding chains and "linker" chains. The latter are required for assembly and structural integrity of the protein and are deficient in or lack heme. We have determined the nucleotide sequences of the cDNA and gene for linker chain L1 of the hemoglobin of Lumbricus terrestris. The CDNA-derived amino acid sequence has 225 residues and a calculated molecular mass of 25,847 Da. The chain is 21-28% identical to linker chains of the related annelid Tylorrhynchus heterochaetus and the deep-sea tube worm Lamellibrachia sp. A remarkable feature of the linker chains is a conserved 38-39-residue segment that contains a repeating pattern of cysteine residues: (Cys-X6)-Cys-X5-Cys-X4-Cys. This pattern, not present in any globin chain, corresponds exactly to the cysteine-rich repeats of the ligand binding domains of the low density lipoprotein (LDL) receptors of man and Xenopus laevis. Furthermore, the cysteine-rich segment of linker chain L1 has the sequence Asp-Gly-Ser-Asp-Glu which is characteristic of LDL receptor repeats. Similar cysteine-rich sequences also occur in two other mammalian proteins, complement C9 and renal glycoprotein GP330. The results support the conclusion that the cysteine-rich motif of the LDL receptor and annelid Hbs is a multipurpose protein-binding unit of ancient origin which has been incorporated into diverse unrelated proteins, presumably by the process of exon shuffling.

Many extracellular Hbs of annelids are composed of two kinds of heme-containing subunits, a disulfide-linked trimer of three different polypeptide chains (17-18 kDa) and a 16-kDa monomer (for review, see Ref. 2) together with additional heme-deficient chains (25-32 kDa) that appear to be required as "linkers" in the assembly of the heme-containing subunits.

The determination of the amino acid sequences of linker chains from the Hbs of the polychaete Tylorrhynchus heterochaetus (8) and the deep-sea tube worm Lamellibrachia sp. (9) has shown that these 25-28 kDa chains are all homologous but are either unrelated or very distantly related to the heme-binding chains. The coding region of the gene for heme-binding chain c of Lumbricus Hb has exactly the same two-intron, three-exon organization found in the genes for vertebrate globins (10, 11). Examination of the linker sequences initially suggested (8, 9) that they might have evolved by fusion of two genes for heme-binding chains to form a gene for a two-domain chain followed by loss of the first exon in the first domain and the last exon of the second domain. We have investigated the organization of a gene for a linker chain in Lumbricus Hb to test this hypothesis.

MATERIALS AND METHODS

Amplification of CDNA of Lumbricus Chain L1—Poly(A)* RNA, prepared previously (11), was used as template to synthesize cDNA with oligo(dT) as primer. The cDNA corresponding to chain L1* was amplified from single-stranded cDNA by the polymerase chain reaction (PCR) using oligo(dT) with an adaptor that included an XbaI site (27-mer, Promega, Madison, WI) and a 2,048-fold redundant oligomer.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank®/EMBL Data Bank with accession number(s) L12688 and L12689.

† Supported by a grant-in-aid for scientific research from the Ministry of Education, Science and Culture of Japan. Present address: Dept. of Biology, Faculty of Science, Kochi University, Kochi 780, Japan.

‡ To whom correspondence and reprint requests should be sent.

\* This work was supported in part by National Institutes of Health Grants GM28410 and GM53547, National Science Foundation Grant DMB 88-10828, and Welch Foundation Grant F-0213. A preliminary account of this work has been presented (1). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

\( X \) stands for all four nucleotides. The reaction mixture (100 \( \mu \)l) contained 2.5 units of Taq polymerase (Perkin-Elmer-Cetus Instruments), 0.3 \( \mu \)g of oligo(dT), 10 \( \mu \)g of redundant oligomer, a 0.2 mm concentration of each of the four deoxynucleotide triphosphates, 40 ng of single-stranded cDNA, and 10 \( \mu \)g of bovine serum albumin, 50 mm KCl, 10 mm Tris-HCl (pH 8.5 measured at \( \sim 22^\circ \)C), 1.5 mm MgCl2, and 3 mm dithiothreitol. Since the enthalpy of ionization is 47.56 kcal/mol (15), the pH at 72 °C should be \( \sim 7.1-7.2 \). The sample was boiled for 2 min before the addition of enzyme. The cDNA was amplified for 30 cycles (1.5 min at 94 °C, 1.5 min at 50 °C, and 2 and 3 min at 72 °C for cycles 1-20 and 21-30, respectively) with a thermal

1 The major structural linker chains will be designated L1, L2, and L3 to avoid the implication in previous nomenclature (12) that they are dimers (D1A, D1B, D2) and to conform to the nomenclature initially suggested (8, 9) that they might have evolved by fusion of two genes for heme-binding chains to form a gene for a two-domain chain followed by loss of the first exon in the first domain and the last exon of the second domain. We have investigated the organization of a gene for a linker chain in Lumbricus Hb to test this hypothesis.

3 The abbreviations used are: PCR, polymerase chain reaction; kbp, kilobase pairs; bp, base pairs; LDL, low density lipoprotein.
cycler (MJ Research, Waltham, MA). Electrophoresis of the amplified products in 0.7% agarose gel in 0.04 M Tris, 0.02 M acetic acid, 0.01 M EDTA gave three major bands: 1.4, 0.6, and 0.4 kbp. These products were isolated from the gel, subcloned into the SmaI site of pUC18, and sequenced by the dideoxy method using Sequenase (U. S. Biochemical Corp.). The 1.4-kbp product was found to be the target L1 cDNA.

The 5' flanking sequence was obtained as follows. Single-stranded cDNA was synthesized from poly(A)+ RNA by primer extension using the nonredundant oligomer 5'-ACGTGGGTGATGTTGAGACTG-3' (24-mer), which corresponds to the complementary sequence of the cDNA in positions 342-365 (Fig. 1). A poly(A) tail was then added to the 3' end of the single-stranded cDNA by terminal transferase (Stratagene, San Diego, CA). This material was used to amplify the 5' region of L1 cDNA by PCR under conditions similar to those described above. The primers used (1 pmol each) are oligo(dT) with adaptor for the XbaI site and the oligomer 5'-CTTCCTTGCGACGTTAGTCGTATGTTGAGACTG-3' (26-mer) (position 101-126, Fig. 1). The products (100-150 bp) were subcloned in PCR-1000 (TA cloning system, Invitrogen, San Diego, CA) and sequenced.

Screening of the λgt10 cDNA Library—The λgt10 cDNA library of Lumbricus terrestris, constructed earlier (11), was screened with L1 clone C26. The insert of one positive clone was cut out by digestion with EcoRI purified by agarose gel electrophoresis, subcloned in the EcoRl site of pUC18, and sequenced.

Amplification of Genomic DNA of Lumbricus L1 Chain—The genomic DNA used is the same sample as in previous work (11). For PCR, 2 µg of genomic DNA and nonredundant primers (1 pmol each)

CA GGC ATA ATG GTC AAC ATG TGG TAC GTC CTA GGC CCT ATG CCT GTC GTC GCG

AC GCC ATC ATC GGC GCT AGC GAT CCC TAC CAG GAG GCG TTT CAG TAC GTC GTT RAG AAC CAG AAC

AGA GAC DPY QEB RSP FQY LVY XNY QN

CTG CAC ATC GAC TAT CAG GAG AAC AAG CAG CTG LHI DYL AKL RKL HNL DIFE EY NKL

ACC CAC GAC GTT GAC AAG AAC ACC ATT CAC GCT CAG ATG GCT GCT ATT ACC AAC CTA GAA

THDVDKKKTIRQLKARISNL

QAG CAC CAC TGC GAC GAG CAT GAC TGG GCA TGT GAG GAA CTA CCA GAG GAC ATC AAC

EHHCDEFSECRGDDVPCE

QAC ATG CTC TCC TTT GGA GAA AAG AGC TGC AGG GAC GAG AAC GAC GAC GCG CAA

DLFLCDFGEGKDCRDSDEDFE

ACA TGC AGT CTC ACC ATC ACC GAC TGC GCC AGC ACC TCA TAC ACC GCC TGC ACC TGC AGC

TC S LNI T HGV O ST YTG L TAT W T

AGG TGC GAG GAC ATC AAC CTT GAC CAT GCC ATC GTC ACC ACC TTC ACC GCC CAC CCG AAA

SC E D L N P H N A I V ITITA A H R K

TCC TTC TCC CCG AAC GCT GTC TGG CTT CGG GCC ACC CTC TAC TAC GAG TGG GAG GAC CAT

SFPPNRWVLFRATALSRYBLDEH

GAC CAC ACG TGC ATC ACC ACC CAG CAT GAG GAT TTT GAC CTA AAC ACC CAA GAA CTC

DHTVSTTTQLRGFPYNFGBKREL

CTT CTC GCT CTT CCG AAA GGT CAG TGC GAG GGT TGC ATC TGC GTC TAC ACC ATC

LLAFLKPGQSEGYGGVICDF

GCC GAT GAT GAC CAC GCC GAC TTC AAG ATC GTC GTC GCG AGC AGT CTG CTC TGC TAC GCA

GGDDADHADCIKIVVPSSLSLFVCA

CAC TTC AAC GCC CAA AGA TAC T AGC AAC GAC ACC AGC CTA TAC AGC TAC ACG CAT CGA GGT

NFNAQYR

GAA ATT TTT TAT GAA GAA GCA AAA TTC CGA GAT CGA ACC ATT GAT CAT CGA ACA ATC TTT

GCA GTC GAA CTT TCC TCT TCT TCG TGT AAA AAT CTT ATC ACA TGG TGC GCC ATC TTT ACC

CAG CAC TGC ACA GAC AAA GTC TCG CCC CTA GCA AAA GAA GCC ACC TAG ACG CGA GAC TTT

GTC TCA TCC GCT GAG AAA TAT ACT TTA AAA GTC ACA CTC TGT CAG TCA ACA ACC GAC

TTT TTA AAG CGG AAT AAC ACT TAA TTA TTT TAA AAC GAT TAA AAG ACA GAC CCT TAT TAT

TAT GCA AGT ACG ACA CGA GGA CGG TCA CTT GAT TAC GAA AGA CAG ACG ATC GTC

GCA GAA ACA TAG AAA TAA GAA CCA AAA CTG ATG AAG TAT CAT TTT TCA TGA ACT GGA

CTG ACG GAT AAA ACA GCA CAA AAA AAC TGC ACA ACA ATG CCG ACA AAA GTA AGC GAA

ACC AAA AAC GAT CTC ATG TAA TTA TCC GAA CTA CAC CCA AAG AGA AGA TCT GAT CTA

GCT CAC GTC GCC TCT CTC CAT TTT GTC CTT CTC ATC TGC AGC TAC GAG TGG GAG GTA TAT TTT

TGC AGA TAC CTC GAA ATC CTA TGA GCT TCA GAA AAA TTA TTA TCC TCA TCA TAT TCA

CTG AGC TGG CGG AAA KAT ACA CTA AGC ACG AAC TAT AAC AAC CAA AAC AAC TAT

FIG. 1. Nucleotide sequence of cDNA for linker chain L1 and the derived amino acid sequence. The sequence was constructed from data of three cDNA clones (5-22, 5-3, and C26) amplified independently. The sequence of nucleotides ~20 to 100 is derived from clones 5-22 and nucleotides 99-1491 from clone C26. The signal sequence is underlined. Arrows mark the positions of splice junctions determined from the combined data of Figs. 1 and 2.
were used. DNA was amplified for 30 cycles (1.5 min at 94 °C, 1.5 min at 55 °C, and 6 min at 72 °C). Other conditions in PCR are the same as described above.

The central region of the L1 gene was amplified using forward and reverse primers, 5'-CCCATAGACACCTAGAGAGAC-3' (position 202-226) in the cDNA (Fig. 1), 24-mer and 5'-GCTTGAGTGTCGAGATCAGCCCGTGA-3' (position 613-636 in the cDNA (Fig. 1), 24-mer). The amplified products were separated by electrophoresis on an agarose gel, and the target DNA was identified with a Southern blot using the cDNA insert of clone C26 as a probe. The fragment (~400 bp) was then isolated from the gel and subcloned into PCR-1000. Three clones were sequenced completely. Subsequent results showed that the forward primer used in this amplification was inappropriate because the genomic DNA corresponding to the primer was split by intron 1 (see Fig. 2). Nevertheless, PCR worked even in this situation.

The 5' region was amplified by PCR using two primers, 5'-ATGTGTTGACGTCTAGCCGCTATG-3' (position 1-24 (Fig. 1), 24-mer) and 5'-TACATCTCCGGGACATTGGC-3' (position 241-264 (Fig. 1), 24-mer). The target product (~2.3 kbp) was identified as described above and subcloned into PCR-1000. One clone (clone 2-26) was sequenced completely, and three clones were partially sequenced.

Finally, the 3' region was amplified by using two primers, 5'-CAGCGGAAATCCATCTCCCTTGGCAGG-3' (position 451-474 (Fig. 1), 24-mer) and 5'-CGTTCGTAGTTCAGTGTACAA-3' (position 1409-1432 (Fig. 1), 24-mer). The target product (~1.7 kbp) was sequenced completely, and three clones were sequenced partially.

A FASTA search (16) for sequences similar to the cysteine-rich segment of chain L1 was performed by the Protein Identification Resource, Washington, D. C.

RESULTS

The cDNA encoding chain L1 of *L. terrestris*, beginning with the position 73, was successfully amplified by PCR using oligo(dT) and a 2,048-fold redundant oligomer as primers. The 1.4-kbp product was subcloned into pUC18, and two independent clones, C26 and C38, were isolated and sequenced completely (Fig. 1). The amplification of the 5' region yielded two additional clones, 5-22 and 5-3. The sequences of these clones differed in 14 positions in the coding region, but these differences were restricted to the third position of a codon and caused no amino acid differences (Table I).

The 5'-flanking region of the cDNA was amplified separately by PCR. The products were subcloned, and clone 5-22 with the longest insert (~150 bp) and another clone 5-3 (~100 bp) were sequenced. The entire cDNA sequence was obtained by combination of the sequences of clones C26 and 5-22 (Fig. 1). The translated amino acid sequence consists of 240 residues including 15 residues of signal peptide. The sequence of the 28 NH2-terminal residues of chain L1 (excluding the signal sequence) corresponded exactly to those determined by direct sequencing (14) except for residue 20. Redetermination of the NH2-terminal sequence confirmed that histidine occupies this position. The earlier identification of leucine in position 20 evidently resulted from carryover from residue 19.

The cDNA library was screened by using the insert of clone C26 as a probe, but no clone with a full-length target insert was found. One clone (AD131) with a ~850-bp insert was isolated and sequenced. The sequence of clone AD131 corresponded to the position 509-1373 in Fig. 1. No nucleotide differences were found between AD131 and C26 (Fig. 1) in the coding region.

The nucleotide sequence of the gene for chain L1 was constructed from three overlapping fragments separately amplified by PCR. No differences in sequence were found in the overlap regions. As shown in Fig. 2, the L1 gene is characterized by three exons and two introns (intron 1, 2,011 bp; intron 2, 717 bp). The sequences around splice junctions, pyrimidine tracts, and possible branch points of introns were confirmed in at least four clones.

The genomic clones yielded five codon differences in addition to those found in the cDNAs of the cDNA (Table I). Codon 111 in genomic clone 41 was CAC (His) where TAC was found in cDNA clones C26 and C38. This is the only amino acid difference found. However, the nucleotide sequence in this region was redetermined with a mixture of 10 other genomic clones, and codon 111 was clearly TAC with only a trace of CAC. We conclude that TAC is the major codon.

DISCUSSION

Protein

Alignment of the amino acid sequence of L1 with those of the three other known sequences of linker chains (Fig. 3) shows that 21 residues are conserved in all four chains; the overall extent of identity is 23-30%. Twelve of the conserved residues (57%) are included in a segment that contains a cysteine-rich motif of the LDL receptor in invertebrates. This is characteristic of the LDL repeats (17,18). Similar cysteine-rich sequences occur in two mammalian proteins, complement C9 (19) and a renal glycoprotein (20). These data, taken together, clearly indicate that these cysteine-rich sequences all have a common ancestry. This is the first occurrence of the cysteine-rich motif of the LDL receptor in invertebrates. The correspondence between the LDL receptor repeats and the cysteine-rich motif of the LDL receptor in invertebrates.
between most of the repeats of the LDL receptor. The conserved cysteine-rich segments of L1 and complement 9 differ from some of the LDL receptor repeats by less than the LDL receptor. The conserved cysteine-rich segments of L1 and complement 9 differ from those determined for chain T1 in cone shell molluscs (27) and other venoms (28-30) which display a variety of connectivities. The cross-linking in these toxins renders them extraordinarily stable; their toxicity can persist in 8 M urea and after exposure to anhydrous formic acid or 1 N HCl (31). Although none of the disulfide connections has been chemically determined in linker L1, their arrangement outside the cysteine-rich segment can be predicted (Fig. 6) from those determined for chain T1 in Tylophorus Hb.3

T. Suzuki and T. Takagi, unpublished results.

3 FIG. 2. Nucleotide sequence of the gene encoding Lumbricus linker chain L1. The sequence was constructed from the data of three genomic clones (2-36, 41, and 10-2) amplified independently. The sequence of nucleotides 1-2251 is derived from clone 2-36, nucleotides 2256-2623 from clone 41, and nucleotides 2486-4166 from clone 10-2. Reverse complement and direct repeats are marked by arrows. The beginning and ending of the long 41-base direct repeats at the end of intron 2 are marked by dots (*).
Gene for Linker Chain of Earthworm Hemoglobin

FIG. 3. Comparison of the amino acid sequence of linker chain L1 of L. terrestris Hb with those of other linkers. LV, linker chain of Hb from the vestimentiferan, Lamellibrachia sp. (9); T1 and T2, linker chains from Hb of T. heterochaetus (8).

FIG. 4. Comparison of the cysteine-rich segments of linker chains with the seven repeats of the ligand binding domain of the LDL receptor of man (17). Symbols are as in Fig. 3. Circled amino acids mark codons that are split by splice junctions in the genes.
apolipoprotein E to the LDL receptor ($K_D \approx 1.2 \times 10^{-10}$ M) (32, 33). The binding depends upon interaction between Asp
and Glu residues of the receptor repeats and a corresponding
set of Arg and Lys residues in a segment of helix 4 of
apolipoprotein E of the cholesterol-transporting particle (33).
Since linker chain L1 has all five of the Asp and Glu residues
that are conserved in the ligand binding repeats of the LDL
receptor, we suggest that a chain of Lumbricus Hb with a set
of positively charged residues exists that is stereochemically
similar to those determined for apolipoprotein E. It is likely
that the conformation of the highly cross-linked cysteine-rich
motif is not appreciably modified by insertion into a different
protein. This conclusion is consistent with the finding that
the hybrid protein formed by insertion of a snake neurotoxin
(also highly cross-linked) into an Escherichia coli phosphatase
has both biological activities (34). The positively charged
groups might occur either in one of the heme-binding chains
(a, b, or c), in the linker chains themselves, or in both. Chain
d is probably not involved because the full hexagonal structure
can apparently form in its absence (4). A possible site for L1
binding is the NH2-terminal helical segment of heme-binding
chain b, which has five positively charged residues in appro-
priate positions (Fig. 7). Presumably, nonelectrostatic inter-
actions between subunits must also be involved because high
salt does not dissociate Lumbricus Hb (12). The LDL motif/
chain b may be one interface, but nonpolar interactions are
likely to be the most important energetically. Linker chains
may have multiple interactions with the heme-binding chains
and are likely also to form linker-linker interactions.

A hydrophilicity plot of linker chain L1 (Fig. 8) comprises
diagram. The dotted connections provided for the
cysteine-rich motif are arbitrary and depict the arrangement found in wheat germ agglutinin and in
several cone shell toxins. The position numbers and symbols for chains are as depicted in Fig. 3.
**Gene for Linker Chain of Earthworm Hemoglobin**

**Apolipoprotein E**

\[
\begin{align*}
\text{Lumbricus } b & \quad \text{HVRASLHRKRLKLDADDLQKH} \\
\text{Chain L1} & \quad \text{HROCGVLEGKVESEWGRAGYSSGHLR} \\
\end{align*}
\]

**FIG. 7.** Comparison of residues 133–158 of helix 4 of apolipoprotein E of man (32, 33) with the NH\(_2\)-terminal segment of chain b of *L. terrestris* Hb (36). This comparison suggests a possible similarity in protein-protein interactions and is not intended to imply any homology.

**FIG. 8.** Hydrophilicity plots of the amino acid sequences of chains L1 and c of *L. terrestris*, determined by the procedure of Kyte and Doolittle (35). Hydrophilicity window = 7.

**FIG. 9.** Possible alignment of the E, F, G, and H helices of chains c and L1 of *L. terrestris*. The assumption is made that residues 106 and 136 of L1 (marked with *) are the distal and proximal histidines. The helix designations are those tentatively assigned to chains a, b, c, and d on the basis of alignment with known structures (36, 37) and model building (38). Residue 70 of chain c corresponds to position 7 of the E helix. The numbering of residues is for chain L1, not the bookkeeping numbers of Fig. 3.

and NO binding kinetics similar to those for intact Hb and to those of subunits containing chains a, b, c, and d (39). This finding suggests that at least one linker chain not only has heme but also may be globin-related. Unfortunately, this fraction has not been characterized as to chain composition.

Comparison of the sequence-derived molecular mass, 25,847 Da, with the mass determined by mass spectrometry (13), 27,728 ± 15 Da, suggests the presence of 1,881 Da of carbohydrate. Both mannose and N-acetylglucosamine have been reported for *Lumbricus* Hb (40, 41). N-Acetylglucosamine is usually bound to an asparagine residue in the sequence Asn-X-Ser/Thr. Although no such sequence is present in *Lumbricus* chains a, b, c, or d (36, 37), one possible binding site is present in L1: Asn-Ile-Thr (residues 103-105).

**Gene**

**Heterogeneity**—A total of 20 nucleotide differences was found in the cDNA and genomic DNA specifying the codons for chain L1 (Table I). Eighteen of the 20 differences involve changes in the third base, and 61% of these involve C ↔ T transitions. We were concerned that the PCR might be responsible for some of these differences. Although we cannot exclude this possibility completely, our reaction conditions are those that have been found to give error rates of \(\lesssim 3 \times 10^{-4}\) or one nucleotide substitution in about 3,300 (42), which is far below the observed frequency of substitution. Comparison of the 3'-noncoding regions showed 30 nucleotide substitutions and 9 deletion/insertion differences (data not shown). This frequency is more than twice that observed in the coding region. No known mechanism exists whereby the Taq polymerase could recognize the codons in such a way as to make 90% of the supposed errors of transcription in the third position of the codons. We therefore believe that most of these differences are not transcriptional errors but reflect the presence of different genes. The more than 2-fold higher number of changes in the 3'-noncoding region supports this conclusion. However, we cannot distinguish between polymorphism and multiple alleles because the DNA was obtained from a large number of worms. A total of at least four genes appears to be required to explain the nucleotide differences in Table I.

The pronounced bias toward C ↔ T transitions (Table I) is about evenly distributed between C → T and T → C in contrast to the strong preference for T → C mutants in the
fidelity of DNA synthesis by Taq polymerase (43). A high proportion of mutants of human Hb appears to result from C → T transitions that are caused by mispairing of cytosine residues in methylated CpG dinucleotides (44). Examination of the L1 cDNA and genomic sequences, however, shows that only one of the 5 C → T transitions is followed by a G. Perhaps the high frequency of C → T transitions in the L1 gene also reflects the presence of methylation “hot spots,” but the mechanism may be slightly different.

**Introns**—The 4,166-bp genomic sequence of chain L1 (Fig. 2) is punctuated by two introns of 2,011 and 717 bp, respectively, which separate three exons of 220+, 524, and 694 bp. The splice junctions are typical of those found in other organisms except that the first intron begins with GC rather than GT. The first splice junction immediately precedes the start codon we do not know whether an additional pre-

4 A preceding intron has been identified in the gene for chain b (R. A. Donahue and A. F. Riggs, unpublished results) and probably occurs also in the gene for chain c (11).

4 The 4,166-bp genomic sequence of chain L1 (Fig. 2) is punctuated by two introns of 2,011 and 717 bp, respectively, which separate three exons of 220+, 524, and 694 bp. The splice junctions are typical of those found in other organisms except that the first intron begins with GC rather than GT. The first splice junction immediately precedes the start codon we do not know whether an additional pre-

5 The first intron begins with GC rather than GT. The first splice junction immediately precedes the start codon we do not know whether an additional pre-

**Acknowledgments**—We are greatly indebted to Siisy M. Jiang for poly(A) + RNA and genomic DNA preparations. We thank Peggy Centillii for construction of the oligomers, Paul Krieger for use of a thermal cycler, the Protein Identification Resource for a data bank search, and Sandy Smith of the University of Texas Protein Sequencing Center for analyses.

**REFERENCES**


