Structural analysis of the gene encoding human gastrin: The large intron contains an Alu sequence
(human Alu sequence/intervening sequences/DNA sequences/S1 nuclease mapping)

RYOJI ITO*, KENZO SATO*, TELSA HELMER*, GILBERT JAY†, AND KAN AGARWAL‡

*Departments of Biochemistry and Chemistry, The University of Chicago, Chicago, IL 60637; and †The Laboratory of Molecular Virology, National Cancer Institute, Bethesda, MD 20205

Communicated by Donald F. Steiner, April 2, 1984

ABSTRACT
We have isolated a human gastrin gene from a genomic library by employing a human gastrin cDNA clone as a hybridization probe. The total length of the gene is approximately 4.0 kilobase pairs, and the gene is separated into three exons and two introns. A 130-base-pair intron interrupts the coding region and a 3.0-kilobase-pair intron is located in the 5′ untranslated region. Nucleotide sequence analysis showed that all of the exon–intron boundaries follow the A-G/G-T consensus sequences. A putative transcription initiation site is assigned to the adenine 60 nucleotides upstream from the exon–intron junction on the basis of S1 nuclease protection mapping. A possible "TATA" equivalent sequence T-T-A-T-A-A is located 28 base pairs upstream from the transcription initiation site. A "CAT box" sequence, C-A-T-T, is located 99 nucleotides upstream of the transcription initiation site. A poly(A)-addition signal, A-U-A-U-A-A, is located 80 base pairs downstream from the termination codon. Comparison of the nucleotide sequences of the human cDNA and the genomic clone revealed that the aspartic acid codon at position 71 of preprogastrin is interrupted by the small intron (130 base pairs). The 5′ region of the large intron contains a sequence of 300 nucleotides that is flanked by 15-nucleotide direct repeats. This sequence exhibits a striking homology to the human Alu-type sequence.

Gastrin, a peptide hormone that exists in several molecular forms and has remarkable structural similarity to another hormone, cholecystokinin (CCK), is one of the most studied gastrointestinal hormones (1). The major physiological actions of gastrin are regulation of gastric acid secretion, antral smooth muscle activity, and growth of the gastrointestinal mucosa. The presence of substantial quantities of gastrin mRNA in gastrinoma tissue compared to normal antral mucosal tissue (2) may be due to either mass concentration of gastrin-producing G cells or an increased level of gastrin gene expression in this tissue. To understand the mechanism underlying this differential regulation of gastrin production, we have begun to study the gastrin mRNA structure and expression of the chromosomal gene as a first step in understanding the regulation of the biosynthesis of gastrin.

Previously, we described the cloning and sequence analysis of porcine gastrin mRNA (3), using the oligodeoxynucleotide probe strategy developed in this laboratory (4, 5). Employing the porcine gastrin cDNA as a probe, we have recently cloned cDNA corresponding to human gastrin mRNA isolated from a pancreatic gastrinoma (6). The human cDNA clone constructed by the procedure described previously (3) was near full length and contained 65 nucleotides in the 5′ untranslated region. This human cDNA clone, pHG156, is longer than the ones described by others (2, 7). With the aid of the human cDNA clone, we have now isolated a human chromosomal gastrin gene. The gene is 4.0 kilobase pairs (kb) long and is divided into three exons by two introns. The larger intron (3 kb) interrupts the 5′ untranslated region and the shorter intron (130 bp) separates the coding region. When this work was nearing completion, Kato et al. reported the structure of the human gastrin gene (8). In this communication, we report that the gastrin gene is 4.0 kb long, instead of 0.7 kb, and contains two introns rather than one intron as reported by Kato et al. (8).

MATERIALS AND METHODS
Enzymes and Reagents. Restriction endonucleases were from New England Biolabs. T4 phage DNA ligase and DNA polymerase I (large fragment) were purchased from Boehringer Mannheim. S1 nuclease was purchased from Bethesda Research Laboratories.

Nucleic Acids, Phage, and Bacterial Strains. M13mp8 and M13mp9 replicative form I DNAs were from New England Biolabs and M13mp10 and M13mp11 RFI DNAs were from P-L Biochemicals. pBR322 DNA was isolated from Escherichia coli LE392 (9). Human gastrin cDNA clone pHG156 DNA (6) was isolated from E. coli HB101. Poly(A)-enriched RNA was isolated from human antral mucosa obtained from total human gastrectomy as described (4). The human genomic library prepared by partial digestion of human fetal liver DNA with Hae III and Alu I and inserted into the DNA of phage Charon 4A through EcoRI linkers was a gift of R. M. Lawn (10).

Screening of Human Genomic Library. Human gastrin cDNA clone, pHG156 (6), was nick-translated to a specific activity of 1.7 × 10⁸ cpm/µg (11) and was used directly as a hybridization probe. Plaques were transferred to nitrocellulose and hybridized with the nick-translated cDNA probe as described by Benton and Davis (12). Hybridization was carried out at 65°C for 16 hr in 0.45 M NaCl/0.0045 M sodium citrate containing 0.1% sodium dodecyl sulfate. Filters were washed six times in 0.30 M NaCl/0.03 M sodium citrate at 65°C and strongly hybridizing plaques were identified by autoradiography. The positive clones were plaque purified and rescreened by the method described above, except that filters were washed in 0.015 M NaCl/0.0015 M sodium citrate/0.1% sodium dodecyl sulfate at 37°C for 1 hr.

Restriction Enzyme Mapping, Southern Hybridization, and Subcloning. Phage DNA was isolated from positive plaques and subjected to restriction endonuclease mapping. The cleaved DNA was separated by electrophoresis on 1% or 0.7% agarose gels with appropriate DNA size markers. For Southern hybridization analysis, the DNA segments were transferred to nitrocellulose filters and hybridized with appropriately labeled probes prepared from human gastrin cDNA clone pHG156. The 5′ probe was the Pst I/Pst I fragment that contained the 5′ untranslated region of 60 nu-
cleotides; the middle probe, the HindIII/HindIII fragment, contained most of the coding sequences; and the 3' probe, the HindIII/Pst I fragment, represented the 3' untranslated region. These fragments were subcloned in M13mp8 and M13mp9 phage DNA by the methods described (13). The appropriate DNA segments from the λ101 genomic clone were subcloned at the EcoRI site of pBR322 and further analyzed by restriction endonuclease mapping and Southern hybridization. The desired fragments of the genomic DNA were further subcloned in M13mp8, M13mp9, M13mp10, and M13mp11 by the method described (14).

DNA Sequence Analysis. All of the restriction fragments, varying in length from 200 to 500 nucleotides, were subcloned in either M13mp8 and M13mp9 or M13mp10 and M13mp11, depending on the restriction site availability in the different M13 phage vector. The desired recombinant phage was selected by hybridization with the appropriately labeled probe (15), and the phage DNA from a positive plaque was isolated. The nucleotide sequence of each recombinant phage was determined by the "dideoxy" method of Sanger et al. (13).

S1 Nuclease Mapping. Single-stranded labeled probes complementary to the mRNA were prepared by the following primer extension–restriction enzyme digestion method. The recombinant M13 phage DNA (2 μg), sequencing primer (0.5 pmol, 15-mer), dCTP, dTTP, dGTP (each 20 μM), and [α-32P]dATP (15 pmol; 3000 Ci/mmol; 1 Ci = 37 GBq) were treated in 12 μl of sequencing buffer containing 1 unit of DNA polymerase I (large fragment) at 23°C for 30 min. Then 0.5 mM unlabeled dATP was added for 2 min at 23°C followed by inactivation of polymerase by heating at 65°C for 10 min. The reaction mixture was adjusted to the salt concentration required for the appropriate restriction endonuclease cleavage. After incubation for 2 hr at 37°C with appropriate restriction endonuclease, the reaction mixture was extracted with an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1, vol/vol) and precipitated with ethanol. Strands were denatured by heating in the presence of dimethyl sulfoxide and then separated by electrophoresis on a 5% acrylamide gel as described (16). The radioactive band was electroeluted and recovered after precipitation with ethanol.

The labeled probe (4×8×106 cpm, 40–80 ng of DNA) was hybridized with human antral poly(A)-RNA (12 μg) in 15 μl of a buffer described by Berk and Sharp (17). The sample was heated at 80°C for 10 min and allowed to hybridize at 50°C overnight. These samples were treated with 200 units of S1 nuclease (Bethesda Research Laboratories) in 200 μl of enzyme buffer at 37°C for 30 min. After digestion, the reaction was stopped by the addition of 1.5 M NaOAc, pH 7.0/100 mM EDTA (40 μl) followed by extraction with phenol/chloroform/isoamyl alcohol. The nucleic acids were precipitated with ethanol in the presence of tRNA at 20 μg/ml. The S1 nuclease-digested samples were analyzed by 8% acrylamide/8 M urea gels along with the appropriate dideoxy sequencing reactions.

RESULTS

Isolation of the Human Gastrin Gene. We have previously described the isolation and sequence analysis of the cDNA corresponding to the mRNA coding for porcine gastric gastrin (3). Recently, we have cloned and determined the nucleotide sequence of the human gastrin cDNAs corresponding to mRNA isolated from pancreatic and liver gastrinoma tissue (6). The pancreatic cDNA clone, pHG156, was employed as a radioactive cDNA probe in the isolation of the human genomic DNA sequence.

When EcoRI digests of high molecular weight total human DNA were fractionated by gel electrophoresis, blot transferred to nitrocellulose paper, and hybridized to 32P-labeled cDNA under standard conditions, hybridization to fragments of 5 and 2.8 kb was observed (data not shown). To isolate these fragments, we screened 7×105 phages of a λ gene library carrying Hae III and Alu I partially digested fragments ranging from 15 to 20 kb. From this library only four phages with gastrin sequences were identified. All of these phages showed an overlapping restriction pattern and therefore originated from a single chromosomal locus. One of these, λ101, which contains 7.0-, 5.0-, and 2.8-kb EcoRI fragments (Fig. 1, lane B), was studied in detail. To determine which of these three fragments contained gastrin sequences, hybridizations of EcoRI-digested λ101 DNA to three different probes of cDNA corresponding to the coding (HindIII/HindIII), 3' untranslated (HindIII/Pst I), and 5' untranslated (Pst I/HindIII) region (Fig. 2B) were performed. Fig. 1 shows that the coding- and 3' untranslated-region-specific probes hybridized to the 5.0-kb fragment (Fig. 1, lanes C and D), while the 5' untranslated-region-specific probe hybridized to only the 2.8-kb fragment (Fig. 1, lane E). These results show that all of the gastrin cDNA hybridization bands that can be found in the total human DNA are also present in λ101, providing additional evidence for a single chromosomal locus for the gastrin gene.

Restriction Endonuclease Map of the Gastrin Gene. To determine the position and transcription orientation of the gene, we constructed a composite restriction map of the 7.8-kb DNA, which is shown in Fig. 2A. To accomplish this, the 5- and 2.8-kb fragments were subcloned in the EcoRI site of pBR322 and the resulting plasmids, p1042 (5 kb) and p235 (2.8 kb) were subjected to multiple restriction endonuclease digestions, followed by hybridization to coding, 5'-, and 3'-region-specific cDNA probes. The 0.38-kb HindIII fragment

![Fig. 1. Restriction endonuclease and hybridization analysis of gastrin DNA clones in bacteriophage λ.](image-url)
Biochemistry: Ito et al.

Fig. 2. Restriction endonuclease map of the human gastrin gene, a cDNA clone, and sequencing strategy for the gene. (A) Composite restriction endonuclease map of the 7.8-kb EcoR1 fragment of DNA that contains the internal EcoR1 (b) site. The dark thick regions identify exons E1, E2, and E3 and the thin line between the exons represents introns. The size of the large intron is indicated by a bidirectional arrow. Only the restriction sites used for subcloning in plasmid or M13 phase and used for sequence analysis are shown. The broken lines with the arrow below the map indicate the restriction fragments used in sequence determinations. The wavy lines above the map indicate the restriction fragments employed in the S1 nuclease protection studies. (B) Restriction map of the human gastrin cDNA clone pPHG156. The terminal Pst I sites are not present in the cDNA; they are given simply to identify the region-specific probes used in these studies.

of p1042 hybridized strongly to the coding-region-specific probe and contained the unique Apa I and Nco I sites of cDNA coding region (Fig. 2B). For location of 3' untranslated sequences, the HindIII-digested p1042 DNA was hybridized to the 3' specific probe. Only the 0.7-kb HindIII fragment hybridized to the 3'-specific probe, suggesting that this fragment contains 3' untranslated sequence. Together, these results show that the coding and 3' untranslated sequences are located on two different HindIII fragments. The orientation of these fragments with respect to one another was established by nucleotide sequence analysis. Since hybridization data (presented in Fig. 1, lane E) showed that the 5'-specific probe hybridized to the 2.8-kb fragment only, the DNA of p235 was digested with various combinations of several restriction endonucleases (e.g., Pst I + BamHI + Nco I, Pst I + Nco I, BamHI + Pst I, and Nco I + EcoRI) and the DNA was hybridized to the 5'-specific probe. Analysis of the hybridization results clearly showed that the Pst I/Nco I fragment contains the 5' untranslated sequences (data not shown).

Since the gastrin gene sequence is located on both the 5.0- and 2.8-kb fragments, it was necessary to orient these fragments with respect to each other. This was accomplished by digesting X101 phage DNA with the restriction enzyme combinations of Sst I + Kpn I, Sst I + Pst I, and Sst I + Nco I, followed by hybridization to the Sst I + EcoRI-specific probe [150 base pairs (bp)]. Hybridization was observed to the Sst I + Nco I (1.8-kb) and Sst I + Kpn I (2.5-kb) fragments (data not shown). These results established the orientation of the 5.0- and 2.8-kb fragments as shown in Fig. 2A.

Detection of Gastrin Introns. S1 nuclease mapping experiments together with nucleotide sequence analysis have shown that the gastrin gene contains two introns. The presence of the small intron 2 (130 bp) was revealed only by nucleotide sequence analysis (Fig. 2A and Fig. 4). The position of the larger intron was determined by S1 nuclease protection studies employing human antral mRNA and the HindIII/Sma I probe (as shown by wavy lines in Fig. 2A). As shown in Fig. 3A, the sizes of the protected DNA fragments were approximately 65 + 4 nucleotides and span the nucleotide sequence A-A-A-G-G-G (Fig. 3A) or antisense strand sequence T-T-T-C-C-C (Fig. 4). Although the S1 nuclease mapping data presented do not precisely localize the intron-exon junction due to heterogeneity in the protected fragments, they do indicate the 10-nucleotide regions where the intron may actually begin. Comparison of the nucleotide sequence of this region of the genomic clone to the cDNA clone suggests that the nucleotide sequence diverges from the cDNA at the 10th nucleotide 5' to the AUG codon as shown by an arrow in Fig. 4. On the basis of the consensus sequences (18), we have assigned intron-exon boundaries, as shown by a wavy line above the sequence and a solid line below the sequence in Fig. 4. As discussed later, the 5' nucleotides T-G-C-A-G (wavy line in Fig. 4) are common to both exon 1 and intron acceptor sequences and for this reason the intron-exon junction begins 5 nucleotides upstream of AUG instead of at the 10th nucleotide, where the sequence divergence is observed.

To determine the intron-exon junction upstream (3.0 kb from acceptor end), hybridization of the Pst I/Nco I frag-

Fig. 3. Mapping of intron-exon junction of transcription initiation of gastrin mRNA by nuclease S1 protection. The subclones and probes used for nuclease S1 mapping are shown in Fig. 2A. The same M13 clone was employed in the preparation of the single-stranded probe and in the dideoxy sequencing reactions. The products of the sequencing reactions were cleaved with the appropriate restriction endonuclease to align the length of the protected fragment(s) with the sequencing bands. The reactions were analyzed on 8% acrylamide/8 M urea gels. (A) The [α-32P]dATP-labeled HindIII/Sma I fragment was hybridized to human gastrin mRNA and treated with nuclelease S1. Lane 1, labeled probe + 12 μg of yeast RNA; lane 2, labeled probe + 12 μg of gastrin mRNA. The lanes marked G, A, T, and C represent the dideoxy sequencing reaction products cleaved by HindIII prior to electrophoresis. The arrow on the left points to the midpoint of the dark region of bands, which corresponds to the nucleotide indicated. (B) The [α-32P]dATP-labeled Pst I/Nco I fragment was hybridized to gastrin mRNA and treated with nuclelease S1. Lane 3, labeled probe + 12 μg of yeast RNA; lane 4, labeled probe + 12 μg of human gastrin mRNA. The sequencing reaction products were cleaved with Pst I prior to electrophoresis. The arrow on the right, along with the nucleotide sequence, points to the band and the nucleotide around which initiation of transcription may begin.
The nucleotide sequence of the human gastrin gene and the amino acid sequence of the preprogastrin. The nucleotide sequence of the antisense strand is shown. The sequences of the coding regions are shown in blocks of three nucleotides and the intron and noncoding sequences are presented in blocks of 10 nucleotides. The amino acid sequences coded by the exons are given above the nucleotide sequence. The exons were identified by sequence comparison of the genomic and cDNA clones and by S1 nuclease protection experiments. The positions of intron and exon junctions are indicated by \( \gamma \). The region of intron not sequenced is indicated by a number followed by a dashed line between the determined sequence. The transcriptional and poly(A)-addition signals, "TATA\(^{+}\), "CAT\(^{+}\), and "AUAUAA\(^{+}\), are shown in boxes. The proposed position of the capped nucleotide is indicated by a star. The nucleotide sequence common to the ends of exon 1 and intron 1 is shown by a wavy line. The nucleotides downstream of the AUG are positively numbered, while the nucleotides upstream, corresponding to untranslated region, are negatively numbered. The poly(A) regions and the 15-nucleotide direct repeats of the human \( Alu \) sequence are shown by broken lines and by horizontal arrows, respectively.

Fig. 4. Nucleotide sequence of the human gastrin gene and the amino acid sequence of the preprogastrin. The nucleotide sequence of the antisense strand is shown. The sequences of the coding regions are shown in blocks of three nucleotides and the intron and noncoding sequences are presented in blocks of 10 nucleotides. The amino acid sequences coded by the exons are given above the nucleotide sequence. The exons were identified by sequence comparison of the genomic and cDNA clones and by S1 nuclease protection experiments. The positions of intron and exon junctions are indicated by \( \gamma \). The region of intron not sequenced is indicated by a number followed by a dashed line between the determined sequence. The transcriptional and poly(A)-addition signals, "TATA\(^{+}\), "CAT\(^{+}\), and "AUAUAA\(^{+}\), are shown in boxes. The proposed position of the capped nucleotide is indicated by a star. The nucleotide sequence common to the ends of exon 1 and intron 1 is shown by a wavy line. The nucleotides downstream of the AUG are positively numbered, while the nucleotides upstream, corresponding to untranslated region, are negatively numbered. The poly(A) regions and the 15-nucleotide direct repeats of the human \( Alu \) sequence are shown by broken lines and by horizontal arrows, respectively.

Discussion

Our results provide evidence that gastrin is encoded by a single-copy gene in the human genome. This gene is \( 4 \) kb in length and shows a simple organization consisting of two introns and three exons encoding the 475 nucleotides that correspond to the unique sequence of the human gastrin mRNA. When our studies on the structural organization of the gastrin gene were almost complete, Kato et al. (8) reported cloning of this human gastrin gene. Comparison of our findings with those reported by these authors revealed significant differences as well as some agreements. In agreement is the finding that the gastrin coding region is interrupt-
ed by an intron of 130 bp that divides the aspartic acid codon as shown in Fig. 4. The major disagreement, however, is that our results clearly show that the gastrin gene is 4.0 kb long, while Kato et al. claim it to be only 0.7 kb long. The S1 nuclease mapping data that led those authors to conclude the absence of an intron in the 5' untranslated region do not agree with our S1 nuclease mapping results shown in Fig. 3A. The results of S1 nuclease mapping experiments performed here without a doubt show that the size of the protected DNA is 65 ± 4 nucleotides and not 175 nucleotides as reported by Kato et al. Our conclusion that the gastrin gene is 4.0 kb long and contains an intron in the 5' untranslated region was arrived at on the basis of several lines of evidence: (i) The presence of a Pst I site in the human cDNA clone near AUG and its absence in the genomic clone suggest a change in nucleotide sequence in this region. Comparison of the nucleotide sequence of the genomic clone with the cDNA clone clearly shows a change in sequence around the Pst I site. (ii) The S1 nuclease protection studies unambiguously established the size of the protected DNA to be 65 ± 4 nucleotides and thus strongly suggest that the remaining 60 nucleotides of the 5' untranslated sequence must be noncontiguous and located upstream of AUG. (iii) The 5' untranslated sequences of the cDNA and the sequence 5' to the Pst I site located 3.0 kb upstream from the translation initiation site show perfect homology. Clearly then, the gastrin gene is composed of two introns and three exons instead of only two exons and one intron as demonstrated by Kato et al. (8). The location of exon 1 in our studies was facilitated by the availability of a full-length human gastrin cDNA clone (6). The human gastrin cDNA clone sequence reported by Kato et al. (7) and Boel et al. (2) lacks a major portion of the 5' untranslated sequence, and therefore this sequence was not available to these authors (Kato et al.) for comparison and generation of 5'-specific probes. However, our previously reported porcine gastrin sequence, which clearly contained 61 nucleotides corresponding to the 5' untranslated region, was not taken into consideration by these authors in reaching the conclusions reported.

The notable finding of the present work is the presence of an Alu-type repeat element in the first intervening sequence. The features of this sequence are the presence of two short poly(A) sequences and a 15-nucleotide direct repeat with one mismatch. Such features have been found in a number of repetitive sequences that have now been located, not only in the flanking sequences of known genes but also within their intervening sequences (22, 23). In vitro and in vivo studies have previously shown the ability of specific Alu sequences to be transcribed, and transcripts of some of these Alus repeats are present in the small nuclear RNA population (24). On the basis of these observations made in other genes, it is tempting to speculate that the gastrin gene Alu sequence is also transcribed in vivo and may be present in the small nuclear RNA populations. Two possible roles for such transcripts have been proposed: (i) these transcripts could be involved in RNA processing; (ii) these transcripts could act as DNA replication primers. Although some evidence is available to support each of these roles, extensive studies have yet to be made to precisely define the role of Alu-type transcripts.

While this paper was being reviewed, Wiborg et al. (25) reported the structure of the human gastrin gene. Although the intron/exon arrangement and the sequence of the promoter region are in agreement with our findings, some differences exist. The notable difference, however, is our finding of a human Alu sequence in the large intron that is flanked by 15-nucleotide direct repeats.

This work was inspired by Prof. R. A. Gregory, whom we honor on his 70th birthday. This work was supported in part by U.S. Public Health Service Grant AM-21901.