Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 1067–1069, February 1984
Biochemistry

Structure of a human gastrin gene

(hormone gene/DNA sequence analysis/intervening sequence/polyadenylylation site)

OVE WIBORG*, LARS BERGLUND*, ESPER BOEL†, FANNY NORRIS†, KJELD NORRIS†, JENS F. REHFELD‡, KJELD A. MARCKER*, AND JENS VUUST*

*Department of Molecular Biology and Plant Physiology, University of Aarhus, DK-8000 Århus C, Denmark; †Laboratory of Genetics, Novo Research Institute, DK-2880 Bagsvaerd, Denmark; and ‡University Department of Clinical Chemistry, Rigshospitalet, DK-2100 Copenhagen Ø, Denmark

Communicated by Diter von Wettstein, November 7, 1983

ABSTRACT **A gastrin gene was isolated from a genomic library of human DNA. The human gastrin gene is about 4100 base pairs long and contains two intervening sequences. Thus, a 3500-base-pair intervening sequence is located 5 base pairs proximal to the ATG initiator codon, while a 129-base-pair intervening sequence separates the region coding for the principal hormonal form of gastrin, the heptadecapeptide, from the region coding for the major amino-terminal portion of the gastrin precursor. The 5′ flanking region of the gene contains the conserved sequences, T-A-T-A-A and G-A-C-T-C-A-T-A-T, in positions similar to those of other eukaryotic genes.**

The peptide hormone gastrin is the major regulator of gastric acid secretion and growth of the gastrointestinal mucosa. Originally gastrin was isolated from antral mucosa as a heptadecapeptide amide, the principal hormonal form (1, 2), but subsequently longer and shorter forms have been identified (3–6). The various forms of gastrin arise by post-translational processing of a common large precursor (7, 8). Determination of the nucleotide sequences of porcine (9) and human (10) gastrin cDNAs showed that the precursor form in humans is 101 amino acids long, while the corresponding porcine sequence is 104 amino acids in length.

For the purpose of further elucidating the regulation of the synthesis of this hormone, a gastrin gene was isolated from a human genomic library. In this paper, we report the nucleotide sequence of this gene.

MATERIALS AND METHODS

Enzymes and Reagents. Restriction endonucleases were purchased from New England BioLabs. Polynucleotide kinase (ATP:5′-dephosphopolynucleotide 5′-phosphotransferase, EC 2.7.1.78; from *Escherichia coli* strain B infected with phage T4) and DNA polymerase I "Klenow fragment" (EC 2.7.7.7; from *E. coli*) was obtained from P-L Biochemicals. Reverse transcriptase (RNA-dependent DNA polymerase, EC 2.7.4.9; from avian myeloblastosis virus) was purchased from J. W. Beard (Life Sciences, St. Petersburg, FL). [γ-^{32}P]ATP (carrier free) and [α-^{32}P]dATP (2600 Ci/mmol; 1 Ci = 3.7×10^{10} Bq) were from ICN.

Isolation of Human Genomic Gastrin DNA. A human genomic library was kindly donated by T. Maniatis. This library was constructed from a limited *Alu* I/*Hae* III digest of human fetal liver DNA that was subsequently cloned in the bacteriophage Charon 4A by using *Eco*RI linkers (11). Screening of 8×10^5 phages was carried out by *in situ* hybridization essentially as described by Benton and Davis (12). The hybridization probe was nick-translated (13) plasmid pHG529 (10), which contains the entire coding sequence of human gastrin cDNA. Phage DNA prepared from positive plaques was digested with appropriate restriction endonucleases, and the resulting fragments were characterized by Southern blotting analysis (14).

Oligodeoxyribonucleotide Synthesis. A pentadecamer, d(C-G-C-T-G-C-A-T-C-T-C-G-T-C-T), complementary to the sequence surrounding the initiation codon AUG in human gastrin mRNA, was synthesized by the triester method (15). Labeling of this oligonucleotide at the 5′ end with T4 polynucleotide kinase followed a published protocol (10). With this pentadecamer as a primer, construction of single-stranded cDNA copies of the 5′ end of human gastrin mRNA was carried out as described (10).

DNA Sequence Analysis. In most instances, the dideoxy chain termination method of Sanger *et al.* (16) was used as described (17). In some cases, restriction fragments were 3′-end-labeled with [α-^{32}P]dATP and DNA polymerase I, and the DNA sequence was determined by the chemical cleavage procedure of Maxam and Gilbert (18). The latter procedure also was used for sequence determination of 5′-end-labeled, single-stranded gastrin cDNA.

RESULTS

Isolation and Analysis of Human Genomic Gastrin DNA. Screening of a human genomic phage library (11) revealed 30 plaques hybridizing to cloned human gastrin cDNA (pHG529) (10). After purification of the plaques, DNA was isolated from all 30 clones and analyzed by the Southern blotting procedure after digestion with various restriction endonucleases. For the initial analyses, plasmid pHG529 was used as hybridization probe. However, this plasmid does not contain more than eight nucleotides of the 5′ noncoding region (10). In order to get a probe specific for this particular region, a single-stranded human gastrin cDNA was synthesized and labeled with [α-^{32}P]dATP. mRNA isolated from a human gastrinoma was used as template, and the pentadecanucleotide d(C-G-C-T-G-C-A-T-C-T-C-G-T-C-T), complementary to the sequence around the initiating AUG of the gastrin mRNA, was used as primer (10).

Four different genomic clones were identified; λHG1, λHG2, λHG3, and λHG4 (Fig. 1). The isolated 30 clones were identical to any of these four. Furthermore, the four different clones represent overlapping fragments of the same region of the human genome. Fig. 1 shows a physical map of this region, covering a total distance of approximately 19,500 base pairs.

DNA Sequence Analysis. Fig. 1 demonstrates the strategy for sequence analysis used to establish the nucleotide sequence of the DNA coding for the entire human gastrin mRNA and about 200 nucleotides of the 5′ flanking region. The human gastrin gene covers about 4100 base pairs of DNA and contains two introns, one of which is about 3500 base pairs in length, while the other one is 129 base pairs long. We have not determined the sequence of the central portion of the 3500-base-pair intron.

For an exact identification of the 5′ end of the gastrin gene, we determined the sequence of the 5′ noncoding part of the gastrin mRNA by a primer extension procedure as described (10), using the 5′-labeled pentadecanucleotide primer complementary to the region around the translation initiation codon of gastrin mRNA.

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

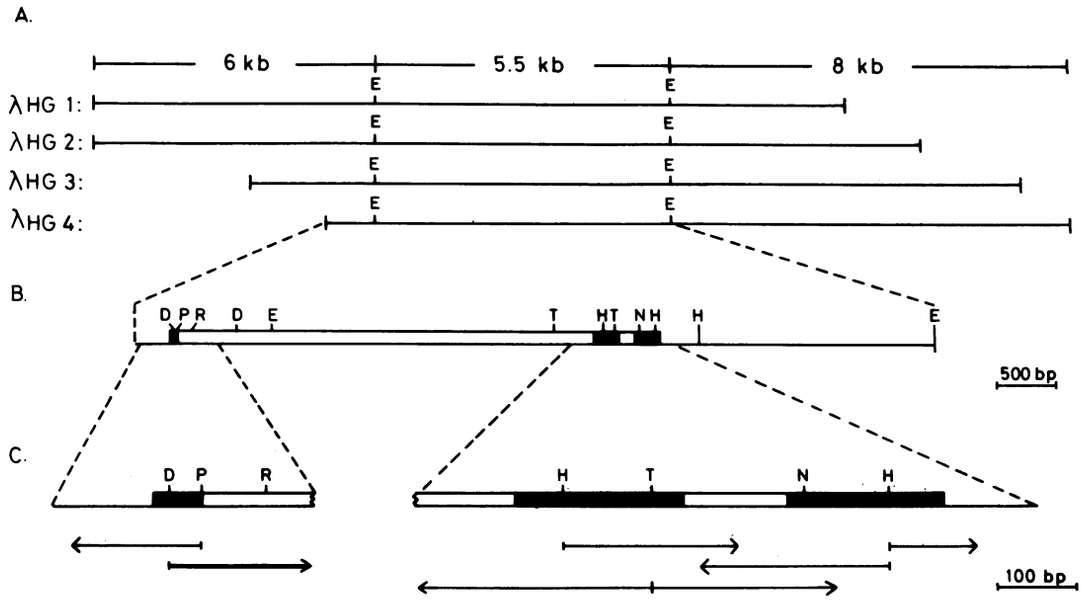

FIG. 1. (*A*) Schematic representation of inserts containing a gastrin gene in four different types of clones found in a human genomic phage library (11). (*B*) Physical map of the human gastrin gene (in the clones indicated in *A*). The boxed area indicates the transcribed region of the gastrin gene. ■, Exons; □, introns. (*C*) The strategy for determining the nucleotide sequence of the gastrin gene. The extent and direction of each sequence reading are indicated by horizontal arrows. Thin arrows indicate sequences determined by the dideoxy technique for sequence determination (16). The thick arrow indicates sequences determined by the chemical degradation procedure (18). E, *Eco*RI; D, *Dde* I; H, *Hin*dIII; N, *Nco* I; P, *Pst* I; R, *Rsa* I; T, *Taq* I.

```
Human                                  -3750                                  -3720                                  -3690
genomic DNA : CCCACCCATCCTCTCGCCTG [GACTCATAT] GGCAGGGTAGGGGCGGGTGGGGGGACAGTTGGGAGGGACCTTGAGGGCTT [TATAA] GGCAGGCCTGGAG

        Cap site                               -3630                                  ↱Intron I                              -3570
CATCAAGCAGAGCAGAGACCTGAGAGGCACCAGGCCCAGCCGTGGCACCACACACCTCCCAGCTCTGCAG [GT] GAGAAAACCCAGGAGGAGAGGGGAGAGGCTAGGAAGTGGGT
        --CGCAGAGACCUGAGAGGCACCAGGCCCAGCCGUGGCACCACACACCUCCCAGCUCUGCAG
          Human mRNA 5' end
                          -3540                                  -3510                                  -3480
TGACAGGTCCTCTCCCCCATCAAGGTACCAGGCCACTGGCCAGAGTCTGGGGCTCACCCCTTGGGGTCTCCAGAGCTGGGACCCTTTCTTTATCCCAGGATGGAACTAGGTCTTG

                                -3420                                  -3390                                                  -120
GCTCCAGTACCTACCCTGGTATTCCCAACCTTGCCTTCCACCTGCCCTTCTGCCGCAGCCGGGG- ( 3250 nucleotides ) -AAAAAAAAAATGAAAGAATTGCCACAC

                     -90                                    -60                                    -30                  Intron I↰
CTCATCAGCAGGTATTAGGCCTGGAGGCCTCTAGGTTCAGTCCCCTGCCTCTGGGCCTCTGTGGGGACAGCCTCACCCTTAAGCTAGTCCCTTCTCCCCTTTGC [AG] ACGAG

1                                      30                                     60
ATG CAG CGA CTA TGT GTG TAT GTG CTG ATC TTT GCA GCG GCT CTG GCC GCC TTC TCT GAA GCT TGT TGG AAG CCC CGC TCC CAG CAG
Met Gln Arg Leu Cys Val Tyr Val Leu Ile Phe Ala Leu Ala Leu Ala Ala Phe Ser Glu Ala Ser Trp Lys Pro Arg Ser Gln Gln
Human protein
 90                                    120                                    150
CCA GAT GCA CCC TTA GGT ACA GGG GCC AAC AGG GAC CTG GAG CTA CCC TGG CTG GAG CAG CAG GGC CCA GCC TCT CAT CAT CGA AGG
Pro Asp Ala Pro Leu Gly Thr Gly Ala Asn Arg Asp Leu Glu Leu Pro Trp Leu Glu Gln Gln Gly Pro Ala Ser His His Arg Arg

    180                                    210  ↱Intron II                240                                    270
CAG CTG GGA CCC CAG GGT CCC CCA CAC CTC GTG GCA G [GT] AGGAGCTGCTGACTGCCCTGCTTGCCTCACTTGGCCAGGTTTGGCAAGGTCTCCCCAGACTG
Gln Leu Gly Pro Gln Gly Pro Pro His Leu Val Ala Asp

                       300                                    330 Intron II↰                                 360
GCTCTGACTTCAGTTCCTGGAAGGTAGGCATCCTTCCCCCATTCTCGCCTCTCTCACCTCCTC [AG] AC CCG TCC AAG AAG CAG GGA CCA TGG CTG GAG GAA
                                                                        Pro Ser Lys Lys Gln Gly Pro Trp Leu Glu Glu

                390                                    420                                    450
GAA GAA GAA GCC TAT GGA TGG ATG GAC TGG GGC CGC CGC AGT GCT GAG GAT GAG AAC TAA CAATCCTAGAACCAAGCTTCAGAGCCTAGCCACCT
Glu Glu Glu Ala Tyr Gly Trp Met Asp Phe Gly Arg Arg Ser Ala Glu Asp Glu Asn xxx

       480                                    510                                    540
CCCACCCCACTTCAGCCCTGTCCCCTGAAAAACTGATCAAA [AATAAA] CTAGTTTCCAGTGGATCAATGGACTGTGT
                                                  pHG315:GG-poly(A)
                                                  pHG168:GGAT-poly(A)
                                                  pHG529:GGATC-poly(A)
```

FIG. 2. The nucleotide sequence of a human gastrin gene. The sequence of the 5′ noncoding part of the human gastrin mRNA, as determined by primer extension, is also shown. The amino acid sequence of the human gastrin precursor (10) is indicated. In addition, the nucleotide sequences of the poly(A) addition sites of three different cDNA clones are shown. In the genomic sequence, the putative promotor sequences and the presumed polyadenylylation signal as well as the splice donor and acceptor sequences are boxed. Because of uncertainty with regard to the exact length of intron I, the nucleotides are numbered from the A of the initiator codon ATG. Negative numbers are used in the 5′ direction. (See *Note Added in Proof.*)

The nucleotide sequence of the human gastrin gene and that of the 5′ noncoding part of the human gastrin mRNA is shown in Fig. 2. It was not possible to read the first couple of nucleotides of the 5′ end of the mRNA sequence. The first nucleotide of the sequence shown (C) differs from the nucleotide in the corresponding position of the genomic sequence (A). This may be a true difference or it may be due to misreading by the reverse transcriptase at the 5′ end of the mRNA.

DISCUSSION

The sequence of the chromosomal gastrin gene derived from normal human fetal liver DNA (11) is in complete agreement with the corresponding cDNA sequence. The cDNA was transcribed from mRNA that was obtained from a human pancreatic gastrin-producing tumor, a gastrinoma; therefore, it is concluded that this highly malignant tumor, which contains high concentrations of gastrin, is expressing a normal gastrin gene.

The transcriptional initiation site, the cap site, was located in the following way. The cDNA primed with the pentadecanucleotide complementary to the sequence around the initiation codon AUG in human gastrin mRNA was about 70 nucleotides in length. If one assumes that the reverse transcriptase reaction goes to completion, this places the 5′ terminus of the mRNA 67 nucleotides from the initiating AUG. The corresponding position in the genomic DNA is an A that is preceded by a C, in agreement with the cap site consensus sequence. The canonical T-A-T-A-A sequence (19) is present 26 base pairs upstream from the putative cap site. Even further upstream, 86 base pairs from the cap site, the sequence G-A-C-T-C-A-T-A-T constitutes a region reminiscent of sequences found in this position in other eukaryotic genes and may have regulatory significance (19).

The human gastrin gene contains two introns. One of these (intron I) is about 3500 base pairs long and is situated in the 5′ noncoding region, 5 base pairs upstream from the ATG initiation codon. Introns in the 5′ noncoding region are commonly found in genes coding for peptide hormone precursors—e.g., preproparathyroid hormone (20), pro-opiomelanocortin (21–24), preproenkephalin (25), prolactin and growth hormone (26), and preproinsulin (27, 28).

The second intron (intron II) is 129 base pairs long and interrupts codon 71. It is located close to the DNA sequence corresponding to the amino-terminal site of the two basic residues that constitute a processing site for gastrin 17, the principal hormonal form. Thus, intron II separates the principal hormonal function from the rest of the gastrin precursor and, therefore, is another example of an intron separating functional domains (29). It is also worth noting that intron II separates the homologous sequences corresponding to codons 41–54 and 74–87, respectively (10).

The sequences around the exon/intron junction at the 5′ and 3′ ends of both introns are in agreement with the consensus sequences reported for splice donor and acceptor sites (19).

By comparison of the 3′ end of the human gastrin gene with the corresponding cDNA sequence of clone pHG529 (Fig. 2 and ref. 10), the poly(A) addition site is located at nucleotide position 534. However, sequence analysis data of two other cDNA clones (Fig. 2) that were isolated in a previous investigation (10) indicate the presence of poly(A) addition sites at positions 533 and 531, respectively. It is not known whether these results reflect true heterogeneity in gastrin gene polyadenylylation or, rather, constitute an artefact of cDNA construction and cloning.

Note Added in Proof. Regrettably, typing errors have occurred in the nucleotide sequence of the human gastrin gene (Fig. 2). The correct sequences are as follows. (*i*) Nucleotides 37 and 38: CT. (*ii*) Nucleotide 65: C. (*iii*) Nucleotides 404 and 405: TC. (*iv*) From position −70 and upstream:

```
                    −130
(3250 nucleotides) −AAAAAAAAAAGAAAGAATTGCACACTC
   −100                                    −70
ATCAGCAGGTAGAGGCCTGGAGCCACATGGTTCAGTC.
```

We thank Iben Hjort, Marianne Nielsen, and Ole Nymann for skillful technical assistance. This work was supported by grants from NOVO Industri A/S, from P. Carl Petersens Fond, and from the Danish Medical and Natural Science Research Councils.

1. Gregory, R. A. & Tracy, H. J. (1964) *Gut* **5**, 103–117.
2. Bentley, P. H., Kenner, G. W. & Sheppard, R. C. (1966) *Nature (London)* **209**, 583–585.
3. Yalow, R. S. & Berson, S. A. (1970) *Gastroenterology* **58**, 609–615.
4. Rehfeld, J. F. (1972) *Biochim. Biophys. Acta* **285**, 364–372.
5. Rehfeld, J. F. & Stadil, F. (1973) *Gut* **14**, 369–373.
6. Rehfeld, J. F. & Larsson, L.-I. (1979) *Acta Physiol. Scand.* **105**, 117–119.
7. Rehfeld, J. F. & Uvnäs-Wallensten, K. (1978) *J. Physiol. (London)* **283**, 379–396.
8. Dockray, G. J., Vaillant, C. & Hopkins, C. R. (1978) *Nature (London)* **273**, 770–772.
9. Yoo, O. J., Powell, C. T. & Agarwal, K. L. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1049–1053.
10. Boel, E., Vuust, J., Norris, F., Norris, K., Wind, A., Rehfeld, J. F. & Marcker, K. A. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2866–2869.
11. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. & Maniatis, T. (1978) *Cell* **15**, 1157–1174.
12. Benton, W. D. & Davis, R. W. (1977) *Science* **196**, 180–182.
13. Maniatis, T., Jeffrey, A. & Kleid, D. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 1184–1188.
14. Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503–517.
15. Ito, H., Ike, Y., Ikata, S. & Itakura, K. (1982) *Nucleic Acids Res.* **10**, 1755–1769.
16. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143**, 161–178.
17. Wiborg, O., Hyldig-Nielsen, J. J., Jensen, E. Ø., Paludan, K. & Marcker, K. A. (1982) *Nucleic Acids Res.* **10**, 3487–3494.
18. Maxam, A. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
19. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383.
20. Vasicek, T. J., McDevitt, B. E., Freeman, M. W., Fennick, B. J., Hendy, G. N., Potts, J. T., Rich, A. & Kronenberg, H. M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2127–2131.
21. Nakanishi, S., Teranishi, Y., Watanabe, Y., Notake, M., Noda, M., Kakidani, H., Jingami, H. & Numa, S. (1981) *Eur. J. Biochem.* **115**, 429–438.
22. Takahashi, H., Teranishi, Y., Nakanishi, S. & Numa, S. (1981) *FEBS Lett.* **135**, 97–101.
23. Cochet, M., Chang, A. C. Y. & Cohen, S. N. (1982) *Nature (London)* **297**, 335–339.
24. Notake, M., Tobimatsu, T., Watanabe, Y., Takahashi, H., Mishina, M. & Numa, S. (1983) *FEBS Lett.* **156**, 67–71.
25. Noda, M., Teranishi, Y., Takahashi, H., Toyosato, M., Notake, M., Nanishi, S. & Numa, S. (1982) *Nature (London)* **297**, 431–434.
26. Chien, Y.-H. & Thompson, E. B. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 4583–4587.
27. Bell, G. I., Pictet, R. L., Rutter, W. J., Cordell, B., Tischer, E. & Goodman, H. M. (1980) *Nature (London)* **284**, 26–32.
28. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) *Cell* **20**, 555–566.
29. Gilbert, W. (1978) *Nature (London)* **271**, 501.