Structure of the human glucagon gene

James W.White and Grady F.Saunders

Department of Biochemistry and Molecular Biology, The University of Texas System Cancer Center, M.D.Anderson Hospital and Tumor Institute, Houston, TX 77030, USA

Received 5 May 1986; Accepted 20 May 1986

ABSTRACT

A clone containing the complete human glucagon gene was isolated and sequenced. The gene is approximately 9.4 kilobases in length and comprises six exons and five introns. The putative preproglucagon encoded by this gene, 180 amino acids in length and containing glucagon and two glucagon-like peptides, is very similar to that of other mammalian species (greater than 90% amino acid sequence homology). There is 88% nucleotide sequence homology between the proximal 130 base pairs of the 5' flanking regions of the human and rat glucagon genes. These sequences, highly conserved throughout evolution, are likely involved in the regulation of glucagon gene transcription.

INTRODUCTION

Glucagon, a 29 amino acid polypeptide hormone found in the pancreas and gut of various species, is an important regulator of metabolism. It acts to elevate the level of glucose in the blood by stimulating gluconeogenesis, glycogenolysis and the release of glucose by hepatocytes (1). A number of other hormones including secretin (2), vasoactive intestinal peptide (3), gastric inhibitory peptide (4) and growth hormone-releasing hormone (5) are similar in amino acid sequence to glucagon and are considered to belong to the glucagon family of peptides.

Multiple glucagon-like immunoreactive species, ranging in molecular weight from 3500 to 18000 daltons are produced by the islets of Langerhans (6). A 69 amino acid form called glicentin (7), consisting of glucagon with an extension of 32 amino acids at the amino terminus and 8 amino acids at the carboxy end, is found in significant quantities in both gut (8) and pancreas (9). The results of DNA sequence analysis of cDNA clones of glucagon-encoding messenger RNAs of cow (10), hamster (11), and rat (12) suggest a hypothetical mammalian preproglucagon of 180 amino acids. The multiple glucagon-like immunoreactive species found in the pancreas can be accounted for by proteolytic cleavage of preproglucagon at different pairs of basic amino acids (10,11). Mammalian preproglucagon contains, in addition to glicentin, two regions called glucagon-like peptide 1 (GLP-1) and glucagon-like peptide 2 (GLP-2) with 48% and 38% amino acid homology to glucagon, respectively (10-12). Two anglerfish glucagon-encoding cDNAs are similar to the mammalian species, but shorter, lacking the region corresponding to GLP-2

(13,14). Partial DNA sequences of human (15) and rat (16) genomic glucagon clones reveal that the coding regions for glucagon, GLP-1 and GLP-2 are located on separate exons, leading to the speculation that the two glucagon-like peptides arose by exon duplication. In situ hybridization studies place the human gene in band q36-37 on chromosome 2 (17).

We have determined the nucleotide sequence of the entire human glucagon gene, spanning approximately 10 kilobases and comprising 6 exons and 5 introns. The sequence of the gene (including the 5' flanking region and the first exon) as well as several differences with respect to the previously reported (15) clone are presented.

MATERIALS AND METHODS

Screening of a human genomic DNA lambda phage library

A genomic DNA library of human liver DNA cloned into phage lambda Charon 4A was generously provided by T. Maniatis (18). Approximately 10^6 phage were screened using a nick-translated bovine glucagon cDNA probe. Hybridizing clones were isolated by two further rounds of plaque purification. DNA from each of the positive phage clones was isolated as described by Maniatis et al. (19).

Mapping of restriction endonuclease cleavage sites

Positions of restriction endonuclease sites within the phage DNAs were determined by simultaneous digestion with two enzymes followed by size fractionation by electrophoresis through agarose gels. Exons were localized by digestion of the clones with restriction endonucleases followed by electrophoresis, transfer of the DNA fragments to nitrocellulose filters, and hybridization (20) with a labeled bovine glucagon cDNA probe (10).

Subcloning and nucleotide sequence analysis

DNA sequence analysis was performed on one of eight clones (λhGCG1) hybridzing with the labeled bovine glucagon cDNA probe. A region of approximately 10 kilobases of this clone, containing the exons previously located by hybridization studies, was digested with restriction endonucleases, size fractionation on polyacrylamide gels, the bands isolated and subcloned into phage M13 vectors (mp8, mp9, mp10, mp11, mp18 or mp19). The nucleotide sequences of the subcloned fragments were determined by the Sanger dideoxy chain termination method (21) using the modifications of Messing (22). Computer analysis of the sequence data was performed using the Bionet computer resource.

Northern blot analysis

Fifteen micrograms of human pancreatic poly(A^+)RNA was denatured with methylmercuric hydroxide and size fractionated by electrophoresis through a 1.5% agarose gel containing 5mM methylmercuric hydroxide. The RNA was transferred to a nylon membrane and hybridized (23) with a radioactive exon 4 DNA fragment of the human glucagon gene, prepared by subcloning the appropriate region of λhGCGl into phage M13 and labeling with ^{32}P-dCTP by primer extension, in a solution containing 0.75 M NaCl, 0.075 M Na$_3$citrate, 50 mM Tris, pH 7.5, 0.1% (w/v) ficoll (400,000 MW), 0.1% (w/v) polyvinylpyrrolidone (360,000 MW), 0.1% (w/v) bovine serum albumin, 250 μg/ml wheat germ tRNA and 50% (v/v) formamide.

RESULTS AND DISCUSSION

Isolation of the human glucagon gene

The human fetal liver genomic DNA library was screened by hybridizing plaque lifts on nitrocellulose filters with a ^{32}P-dCTP labeled bovine glucagon cDNA clone. Of approximately 10^6 plaques screened, eight hybridized strongly with the probe. Five clones were chosen for further study, four of which had identical restriction endonuclease digestion patterns while the fifth was approximately 2 kb shorter at one end and 2 kb longer at the other end and was missing at least one exon, as determined by Southern blotting to a bovine glucagon cDNA probe.

Nucleotide sequence of the human glucagon gene

The nucleotide sequence of approximately 10 kb of λhGCGl was determined by the Sanger dideoxy chain termination method (Figure 1). Although the sequence of the human glucagon mRNA has not been determined, the similarities among glucagon cDNA clones of other mammals (cow (10), hamster (11), and rat (12)) and segments of the human gene permit the assignment of the most probable positions of intron/exon splice junctions and the 5' and 3' termini. The human glucagon gene has six exons and five introns and spans approximately 9.4 kb (Figure 2). The nucleotides at the extreme 5' and 3' ends of all five introns match those of the consensus splice sequences, GT and AG, respectively. All but 9 nucleotides of the 5' untranslated region of the message are contained in exon 1. Exon 2 represents the signal peptide and part of glicentin related pancreatic peptide (GRPP). Exon 3 encodes the remainder of GRPP plus all of glucagon. Exons 4 and 5 encode GLP-1 and GLP-2 respectively, while exon 6 contains the last 4 nucleotides of the coding region plus all of the 3' untranslated region. The sizes of the introns and exons are collated in Table 1. In the approximately 600 nucleotides of 5' flanking region sequenced, a TATA box was found (from nucleotides -24 to -19), and while no CAAT box was seen the segment from nucleotides -68 to -65 represents the complement of the consensus CAAT sequence. This is the same positioning Graves et al.(24) found for the CAAT region of the herpes simplex virus thymidine kinase gene, which stimulates transcription even though it is on the non-coding strand of the gene, an orientation opposite to that normally seen.

The proposed start of transcription was determined by comparison with the rat

```
gaattcatttattaaaacagaacacataggggtttaatcaatatccttaaatttccacaaacataacat -533
aaataaactccacgttgtgaggaagagaggattttaatacatatgtgttgaatgaatgatcattattta
gataaatgaatgactgaagtgattgttatattcaggtaaattcatcatggctaggtagcaaaccaaagac
ttgtaagaacctcaaatgaggacatgcacaaaacagggatggccatgggctacgtaatttcaaggtcttt
tgtcttcaacgtcaaaattcactttagagaacttaagtgattttcatgcgtgattgaaagtagaaggtgg
atttccaagctgctctctccattcccaaccaaaaaaaaaaaaaaaagatacaagagtgcataaaaagttt
ccaggtctctaaggtctctcacccaatataagcatagaatgcagatgagcaaagtgagtgggagagggaa
gtcatttgtaacaaaaactcattatttacagatgagaaatttatATTGtcagcgtaatatctgtgaggct -43

                                                 1
aaacagagctggagagTATATaaaagcagtgcgccttggtgc  AGAAGTACAGAGCTTAGGACACAGAG

CACATCAAAAGTTCCCAAAGAGGGCTTGCTCTCTCTTCACCTGCTCTGTTCTACAGCACACTACCAGAAG
gtaagatgattataaaattgtaaatcctgtttggcggacagtgaagtatttttaagggatcaaaatatta 166
taaattaaaatgttgttctttcatcttagactttattactaatagtacacagagaggaactgagatggaa
aaggttatatcaaatgcatttacgtgtactttaatatagcgaacggcaaagcgagttggaaaaataatta
tatgcaaaaatataaaacagaaaaaaaacaaagattcaaatcaatgcacttgttataatacttaactgtt
atgagagttgtattttaaaataattggtaacattttgaagaataaatatttttcttggacttgatagatc
tgtatactactttgaacagaaggcgtcttttaaagtaagcagaaatgtgtccattagggagcctataaca
gaaattgctttttaccaattaaaatctctggttttccaaaagagcaaattaaataacatctttcaaatat
tcaaatttcagacatctatcaaaaaatcaaagtccattaaagcaactctttgtaaataaaacatagcgta
cttggtgcagagaacaagtgtcatctatttggggggtttttctttggatgcatctgagtgaagccatacat
taataactatttactatgtattgaggataaatagtaataaaatctaaaatagctaccctctgcaataaaa
ttttaatttgtctttttagattaggctcctagagacaaaaaacaaatttacaaagatctttgatggagta
aatgattaagtggtgtatttttccatacatactgaagacctattacataaaatgagctaaccagtttgaa
aaaggatttatatggactacaatgccaacacatttggggatagaacataatgtcttctttttaggtcta
aggttaatctttctgtaagacttcctacttattattgttaccatactttcctttaaattttagtggaaat
ttgtagttcttacaaaactgcaggctctctttagactacctaaagagggggggaaattaactggaaattat
tttcttattgaataagattttaatactaaaataacataaatctctactttctcagcaattgtattcaaac
aattaactctattcttttttttttcacaattcgatagactttcacaaacaaataagtgagattttaaag
ccagttctttcttcagaagaagtcatttattgttggtaatattattagatacggacgctcactttcttat
gtggtttatggttgttctttcttccatcagctcctgaagtggcaggatacctcctgctctaaatcctcct
caatggaacataataggagatattgcaaaacgcttaggactttgggtggaatgaatttttttttcagtct
tgtgaaagactaaaagcttttacaagaaacttttcaacagtggtataatttttaaaagaagtgcttactt
taatatttaaatattcctgcaagcacggatataacaattgagagttataatcagtagatactacctaaat
atatggcaaggtttccaaagtctgccaaccggttaaactgaagcatagcaaaatgtcaaattggtaattt
gatgttgatgaatgaagaaaggtgaatatatcattcctattatgttcctttgtctaatcatccattcaat
gtacaaaaaacttaatatttcaaagaaaaaagttactgctgcaggtattacagtgtctatgttttagtgc
cagagatagtaaatacagctaatatattaaggaaaagattgatagtttaaaaatgataatatttgtgaaa
tccttattccaaaccaaatgtcatttgctaaaaatattcattcaaattatttgtggataagcatagtcaa
acttggattgaaatgcccgcttctcaaaggaaacaccattgggtttcaaagtagccgagatccatgacaa
agacggacttgacgttcccattcagaaaatagaagcgcttgctctgtgggacattctgtcccacattcat
gttggtcccagtccccgtcgacactgcatcatcacatccttaacatactatggcttcgcagggcatctgc
ttgtgacacagggacaagcaatcaggtgtatcctataaccaagtctatggtgtcaccagcagtgaatgac
atgtcctcatctaagtccaattagaaaacaacctggggactttatgactttagaagcaagctctaattt
ttttttacatagtaatctcttagtttttttagttgtgattgttcacgttgccagaagcattagttcttgtt
gcaaattgcttgtctttaaaatgatttttttcccttttaatttagaacaggtcaaagcaccctaaaatacct
caaatctgttggccgctttataattacaaagtctagaccaagttacactatttaaaatataaatgaataa
tacatctgaaccgacagctggtatgccaagttgtagcctgaggctgctttaagatggttgaaatgcagta
taatgtaaatgcttggaagtgaaggactattaattgaaagccagactgatagtgaggggggtggcatgat
ggataaggtgccattcaggctaaccatatgtagcatatcatctacagtgggtattttaaatgttgattgt
ggacattttaatttccagtctggaaatgagccttctaattattaaacagaagtgtccctacattaatgaa
aaagcttaatggtggaacacatctttgtgtcctttttttgcctccaaaaaatctggggaaataatataac
ccactatttaaaattaagctgaaaatataatcagaataaaagtgataacactagctttttccttctactt
atgatatttatctagtcaaatctaattaatttagcctgacatgtttaaaaatccttgcctgccccccctca 3036

                                      Met Lys Ser Ile Tyr Phe Val Ala
ccctacccccattctgtgttctgacag ACAGCAGAA ATG AAA AGC ATT TAC TTT GTG GCT

Gly Leu Phe Val Met Leu Val Gln Gly Ser Lrp Gln Arg Ser Leu Gln Asp
GGA TTA TTT GTA ATG CTG GTA CAA GGC AGC TGG CAA CGT TCC CTT CAA GAC

Thr Glu Glu Lys Ser Ar
ACA GAG GAG AAA TCC AG gtattaaatccgtagtctcgaactaacatatcaatatggttggaat 3210
aaagcctgtgaaaactatgattagtgaataaggtctcagtaatttagaataaatattctgcacaatgatc
```

```
aaatgtttaaagtatccttgtgataaaagcagactttgttggagtgtagatcagagtttgtgtttttat
aataaaaggggagagaaatgttgaaaaacatgttacaggggaaattatagctccttgaaatctacagatg
atcctatctatcaattcattactcacacactgcataatagattacttcaattattcgcaccaactctttt
ctctttttcctcctttaaaaaatagcattatttcttcctaaatgtaaaagccatacatgttcatgatgg
aaagtatgaaaaatatgcaaaaaatattaagtactcaaaattcctctgtccaaagaaagctattcaaagt
aaacaaatttggtgtattactttcctgtgttttacgtaaactgtacataaatatctcttggctcattata
tggctttgtatcatgcttaatatcttaacattgtattattatacactttattatattattaaaaatatt
tgttaaacatgatatttaatggcagcatactgtttcatctcataaatgtacaatatttgtttagctactt
atttgggaagcttaagaggtcttaaaatttattccacaaagaatgttgcaaacaacattttatatacaca
ctttcctacatttcttactatttttttgcacataggtttctagttgtagaaatactgtgtcaaagggcct
gaatgagtttaaaccttttttaatatcttgccaacttgctttccagaaagttcatacacatgcacactct
catcagcagtatagaaatactcctcataatatcctcctcagcattcaagctttgctaatttcataggtaa
aaatagtatctcatttctgagtattaattacattgactattttcccatgtatttaccatgccaaccaact
gttgattagtggcatagggattattcctgatacagtttcaaattagcaagtatcctttttcttgccttte
tttgaacttctagttgtcttccttccactgctagccaccatataactagataattgtgctgttttcatat
taatatatgcaaaacaaaattgggagcttaaatctgctgcccaaagaaatgatttgggattcttcactct
tttaaaatattcaagtcactactctcttctgtcactattagcataatcatgagtagattaattagtagaa
aaagatttgtttaatcctatttcagaaaaaaagaagtcagttgagaaacatgtttctagaattcagaata
atagcgtatgcaactatactaaaacgaagggattctcatgacaagctaaggaagatctttctaaactacc
tattgaaatactctagatgcctgccttactgttttatatggtcttgtatttgtagtgaagatgcttctc 4680

                                                   g Ser Phe
aagtgagtctactcttgaggagagatttatgttgtaccaatcactgttcttcacag A TCA TTC

Ser Ala Ser Gln Ala Asp Pro Leu Ser Asp Pro Asp Gln Met Asn Glu Asp
TCA GCT TCC CAG GCA GAC CCA CTC AGT GAT CCT GAT CAG ATG AAC GAG GAC

        GLUCAGON
Lys Arg His Ser Gln Gly Thr Phe Thr Ser Asp Tyr Ser Lys Tyr Leu Asp
AAG CGC CAT TCA CAG GGC ACA TTC ACC AGT GAC TAC AGC AAG TAT CTG GAC


Ser Arg Arg Ala Gln Asp Phe Val Gln Trp Leu Met Asn Thr Lys Arg Asn
TCC AGG CGT GCC CAA GAT TTT GTG CAG TGG TTG ATG AAT ACC AAG AGG AAC

Ar
AG  gtaagagtctaagcctggctcaaaacttgcttataaatgtattaaataggtctaaaatttctcttg 4964
acattattaagttcttatgctcgtcaaatagtgcccataagccctgatcattttgaagatgtttagggtt
ggggatgctctccttagtattatgccccgaatgtcttaacccgcacccctctaaaccacttacccaacc
tccagtgtgacatagcctttaaattattttcccagtggtagtcctagcaatatttacttgcttactcatt
ttcccttaaagtgaaagtttactttgtggacctacgaggttaccaatctaatttctaccaaaagataga
gcaaatataacattctttcgtggggttcaaacccaacattaaaattatcccccagccaagggtaaggaat
ggaaacagatgcattccttttcttgtagaataaaatccaagacatgcatttaaaattacagcctagctgc
acagcatagagagagaaggcagcaaataagaacacaagcctatcgtaaggcatacattttaatgttggat
tttctttgacgagtgttgggtttttgctgcatctaggctactggaggaaattatccgcaaggtttcaca
gagatggtggaatttcacaatttttttggagggtggcccacaagaggaagggagttatgcatagaaagta
gctgaagaaatcaaaaaatcaggaaaaggacatagatttgtctggatagactggagaagagagagagtga
gagtaagaaaggagagagagagagagagagagagagagagagagagagagaacaagcaagcaggaggaatag
gagaaacagagttgctaatcaaaggttcttaaagtagagggtacagccacaggaagatggctatcatggt
cagagcagacactgggagaggtccctcttaaggactacagtgaaaggtgagaaatcacaaatttataaaa
atgcaaacaactgtcctcatttctagaatgggaatgccacaacttggtaaatggatgtaaagcaatcct
cacactcaagcagtgttgcctatgtgtttaaaatgaggtgatatggggtaatatgtgccagggagactga
cagcagctgatggctcacctccacggtgctctgcacacacttgatgttgatatggggtcatcttctttgc
tttccttgttactagcagatcacacatttcttaatccattatcaagtggtcctttggtaacaattca
agatggtcctttgctttccctcaaatcaattaactaactctgaccccatacttactaataacaaacagct
gccttgcttgcaagtcacatagtgcagtggatagcctgtatccattgagactgaagtcaaaggtaaggca
gctttaataaccaatgacatacataagcaggggattgcagatgggtatgctgggaaggaggggggagctt
tagcccaccgattgttattagctctcttccaccagtttcaatccagaacattaatgtagcttcacgacaa
atcccctagcaaccctcatctcctaaactccctaagccccttctacatagaatcctgaacccaaagttgc
cactgcttataggtgagaaccatattgccaaagagacagatcttcaatttaactttcacatttctttcag 6574

                          GLUCAGON-LIKE PEPTIDE 1
g Asn Asn Ile Ala Lys Arg His Asp Glu Phe Glu Arg His Ala Glu Gly Thr
G AAT AAC ATT GCC AAA CGT CAC GAT GAA TTT GAG AGA CAT GCT GAA GGG ACC
```

```
Phe Thr Ser Asp Val Ser Ser Tyr Leu Glu Gly Gln Ala Ala Lys Glu Phe
TTT ACC AGT GAT GTA AGT TCT TAT TTG GAA GGC CAA GCT GCC AAG GAA TTC


Ile Ala Trp Leu Val Lys Gly Arg Gly Arg Arg As
ATT GCT TGG CTG GTG AAA GGC CGA GGA AGG CGA GA  gtaagtctgtacattcttattt 6734
gacatttttgccttgatgcagaaaatttaagactacagttatctatatatggatctggattacagaagc
aattagtagtcttgcaaagtaaggaaataattcctattgatgaaaaacagtatataaaagttaaacccat
tttgtttttggtactaagtattaataatagagccaaacaggttacatttgtatccccttatagttgcatt
ataattaggtattaaatctttgccaaaccaggccacctctgagagtaaataggatgtttttaataactac
gctgagacaaatttaaactaggactgttctaggggagctagagataaggaaaaagaagaaacagtgctga
aaacttcaaatatgtaaagaaaacatataatatttaaagcttaactttaaacatttacatatgtgatcaa
catatatttatatattaataaatattcagataaaaatgtgatgaagactgaaagtgaccaaacctaagat
gttgataaaatgatattcaaagtacaaataggtaaaaacatccatatatttaactacatatccaatttg
tatggggctgtcatatagatatctactaaataatctaagttgaaaaacaaccaagaccatcaattacttg
cttagatcttaacacagccaaacagacccctgaaccatctcatttcttccgatttttttggagagatg
aaatatgagagacggagaatttatgttcaactctgatttttaaattagatttaaaacaagcttactgaaa
tttaaagagatctcaagatgaaagagaactagaataatggttggtttttaaaacattaataattaaactt
ataaaaccaatggtaaaatagtttctcactcttgacgatattttgcagtgttttaaagggataccaaaaa
ttctgcaatagtaaaccagtgaaagagaaaaatctaatatagatgaagctttaacctcttaatactgcat
tttgcaaggctgttcctgcaagctctggttttataggatatgatatatttagttgaattacagactaata
atctcaacaatagtttctgtattgtcaatatactaaaatcttcaaaacagcctagaagattgaaaagggc
atgaaattatgcaggcttggtattagatcccagctctgctactttctctttccagtagtcactggtccac
atgggtttttatcaggtatttcacagtacaccctttaaaaacagaatcctttactgtttcctcaagacac
ttgtgcatgttaccagtggtagacaatctgtgatcatttattgaaaatatctaatcaaacgctaattta 8064


                   p Phe Pro Glu Glu Val Ala Ile Val Glu Glu Leu Gly
acacttattttcttag   T TTC CCA GAA GAG GTC GCC ATT GTT GAA GAA CTT GGC

        GLUCAGON-Like PEPTIDE 2
Arg Arg His Ala Asp Gly Ser Phe Ser Asp Glu Met Asn Thr Ile Leu Asp
CGC AGA CAT GCT GAT GGT TCT TTC TCT GAT GAG ATG AAC ACC ATT CTT GAT


Asn Leu Ala Ala Arg Asp Phe Ile Asn Trp Leu Ile Gln Thr Lys Ile Thr
AAT CTT GCC GCC AGG GAC TTT ATA AAC TGG TTG ATT CAG ACC AAA ATC ACT


Asp Ar
GAC AG  gtgactgcttttagttaattctgaaaaccatcaaattctcataagtactgattgtcatact 8286
gctgcaagctgttccccatgtaggggaaagtctataatctcttattatataggaaataattatgcatgtt
taagtcatataagagtacatcattatgtttgttttgtatcattacagtttgtatctatttattcatttat
tccagaaacatttactgagtgtctattagagtcaagaagctagaaacacagatagaaatgaaacatggga
aatgttacatcattcacatattgtcacattaaaaattgctatgttgaaatttcagcacaaggtgaaatgc
agcatatacaatattacatacttaatttaataagaagagtagtgagaactggacaccgaaaaatacttat
cttgatgaatttgatttttaataattattcataattgtttgatagcattgatgagcagactttaggat
aaataatctttaaatgaaaatattttaagagtccaaaggatggagggatttatggacaaagggattaata
aacaactttatcaaaataaatcattgaaatattttctggataagttaatgatatcatcttattataatt 8846


                                  g Lys Oc
atgttaaatcatttttcttttttttaatctctag   G AAA TAA CTATATCACTATTCAAGATCATCTTC
ACAACATCACCTGCTAGCCACGTGGGATGTTTGAAATGTTAAGTCCTGTAAATTTAAGAGGTGTATTCTG
AGGCCACATTGCTTTGCATGCCAATAAATAAATTTTCTTTTAGTGTTGTGTAGCCAAAAATTACAAATGG
AATAAAGTTTTATCAAAATATTGCTAAAATATCAGCTTTAAAATATGAAAGTGCTAGATTCTGTTATTTT
CTTCTTATTTTGGATGAAGTACCCCAACCTGTTTACATTTAGCGATAAAATTATTTTTCTATGATATAAT
TTGTAAATGTAAATTATTCCGATCTGACATATCTGCATTATAATAATAGGAGAATAGAAGAACTGGTAGC
CACAGTGGTGAAATTGGAAAGAGAACTTTCTTCCTGAAACCTTTGTCTTAAAAATACTCAGCTTTCAATG
TATCAAAGATACAATTAAATAAAATTTTCAAGCTTCTTTACCATTGTCTGATTTCTCTTTACGTCCCCCT
TTGTCATGCCCCACTCCTTCCAAGCACCCTGTTTTCTAAGCTGCAGT                        9448
```

Table I
Size of the Transcribed Regions of the Human Glucagon Gene

Exon 1	~96 bp
Intron A	2967 bp
Exon 2	101 bp
Intron B	1572 bp
Exon 3	162 bp
Intron C	1676 bp
Exon 4	138 bp
Intron D	1368 bp
Exon 5	144 bp
Intron E	654 bp
Exon 6	~509 bp

sequence (Figure 3), in which the start of transcription has been identified by S1 nuclease mapping (16). Since the region from -9 to +5 is identical in the two genes, and the TATA boxes are in similar positions, we propose that the transcription initiation sites are most likely equivalent in both genes. There is 88% nucleotide sequence homology between the first 130 nucleotides of the 5' flanking regions of the human and rat glucagon genes, and the 44 nucleotide stretches containing the CAAT sequence from -93 to -50 are identical. This high degree of conservation suggests that these regions are functionally important, perhaps to the transcriptional control of the genes. On the other hand, there are no detectable similarities between the sequences of the introns of the human and rat genes except near the intron/exon boundaries (within 6 to 27 bp, depending upon the boundary). In these short stretches, 72-100% of the nucleotides are conserved between the two genes. Without the sequence of the human glucagon mRNA, it is difficult to define the 3' end of the gene. The poly(A^+) additional signal (AAATAAA) is in a similar position in all the mammalian glucagon cDNA clones (10-12), but the number of trailing nucleotides until the poly(A^+) tail is added is slightly different for each species; 19 for cow (10), 15 for hamster (11) and 9 for rat (12). Due to these similarities and the size of the human glucagon mRNA, we propose that the 3' end of the gene lies in an equivalent region, somewhere around nucleotides 9363-9373.

Comparison of the sequence of λhGCG1 with that of the previously reported human glucagon clone (15) reveals 20 regions containing sequence discrepancies (Table 2). It is possible that the two clones represent different alleles, since both clones were

Figure 1. Nucleotide sequence of the human glucagon gene.
Bold capital letters represent exons. Lower case letters are introns and flanking regions. The TATA box (-24 to -19), poly A addition signal (AAATAAA, 9346 to 9352), and CAAT box complement (-68 to -65) are capitalized and underlined. Numbering begins from the putative cap site.

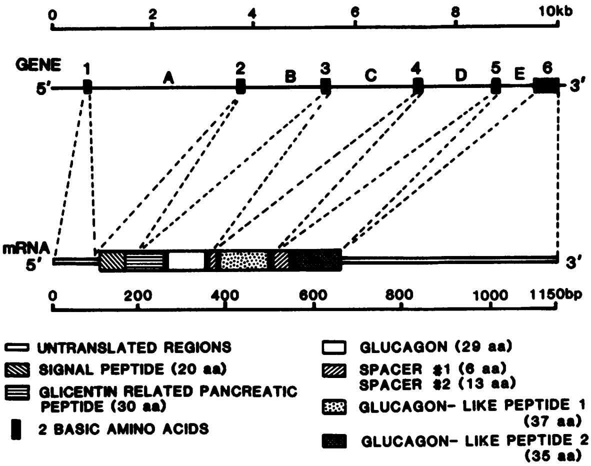

Figure 2. Structure and splicing of the human glucagon gene.
In the diagram of the gene, the filled boxes represent the six exons, while the horizontal lines are the five introns and the 5' and 3' flanking regions. The putative mRNA, obtained by splicing together the exons, and the portions of the preproglucagon molecule represented by the coding region of the mRNA, are shown in the lower part of the diagram.

```
       -120      -110       -100
acgagagtgggcgagtgaaatcatttg-aacaaaa
* * ******* *** *** ******* *******
aagtgagtgggagagggaagtcatttgtaacaaaa
        -120      -110      -100


    -90       -80       -70       -60
ccccattatttacagatgagaaatttatattgtcagcgtaat
 * ***************************************
actcattatttacagatgagaaatttatattgtcagcgtaat
    -90       -80       -70       -60


  -50       -40         -30
atctgcaaggctaaaca--gcttggagacta
*****  **********  *** ***** **
atctgtgaggctaaacagagct-ggagagta
  -50       -40        -30


   -20       -10       -1
tataaaagccacagcaccttggtgcAGAAGGGCAGAGC
***********  ** **************  ******
tataaaagc-agtgcgccttggtgcAGAAGTACAGAGC
  -20        -10       -1
```

Figure 3. Sequence comparison of the 5' flanking regions of human and rat (16) glucagon genes.
The upper sequence is from rat and the lower from human. The flanking regions are represented by lower case letters, while transcribed sequences are in capital letters. Asterisks identify matches between the two sequences. Hyphens show where gaps have been inserted to maximize homology.

Table II
Differences in Glucagon Gene Sequence with that of Bell et al. (15)

Insert G between 3388 and 3389
Delete C at 5610
Change G at 5866 to C
Delete C at 5880
Insert T between 5952 and 5953
Change TG at 5984-5985 to GT
Change CA at 5991-5992 to AC
Delete A at 5996
Insert G between 6041 and 6042
Insert G between 6119 and 6120
Delete GAGAGA from 6310-6315
Change G at 6330 to A
Change A at 6332 to G
Change A at 6705 to G
Delete T at 6773
Delete C at 6803
Delete T at 6809
Change A at 7390 to G
Insert G between 7568 and 7569
Insert A between 8973 and 8974

These changes to the λhGCG1 sequence would generate the sequence of Bell et al. (15). For reference, position 3516 of λhGCG1 corresponds to position 9 of Bell's clone (the first 8 nucleotides of which are probably a cloning artifact).

isolated from the same library and since there is no evidence for multiple glucagon genes in mammals. Alternatively, the differences could be due to cloning artifacts, misread sequencing gels, or typographical errors. The sequence presented here extends that of Bell et al. (15) by 3515 nucleotides at the 5' end (which includes the 5' flank and first exon) and 82 nucleotides at the 3' end.

Size of the human pancreatic glucagon mRNA

Human pancreatic poly(A^+)RNA, size fractionated on a denaturing agarose gel and transferred to nitrocellulose, was hybridized with a ^{32}P-dCTP labeled M13 clone which contained a 580 bp fragment with exon 4 of λhGCG1. The results of this experiment (Figure 4) indicate that there is only one size species of glucagon mRNA in human pancreas and it is about 1250 nucleotides long.

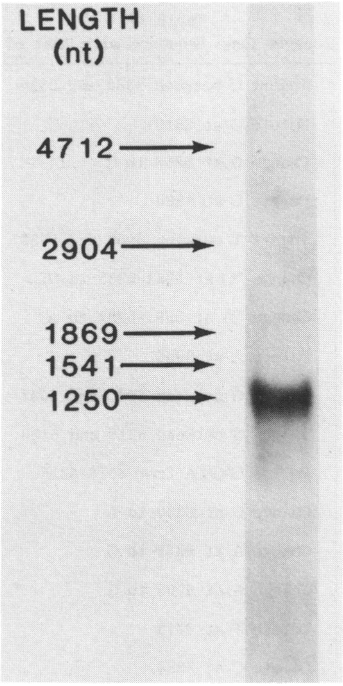

Figure 4. Size of the human glucagon mRNA.
15 µg of human pancreatic poly (A^+)RNA was size fractionated on a 1.5% agarose gel containing 5 mM methylmercuric hydroxide, transferred to a nylon membrane, and hybridized with a ^{32}P-dCTP labeled probe containing exon 4 of the human glucagon gene.

Sequence analysis of the human glucagon gene

Computer analysis of the glucagon gene sequence demonstrates that there are 6 inverted repeats and 3 direct repeats longer than 11 base pairs in this gene. The two longest pairs of direct repeats are 13 bp in length: CACTATTTAAAAT (2522-2534 and 2898-2910) and TTTTATCAAAATA (8783-8795 and 9059-9071). The longest inverted repeat is 14 bp long: GGAACATAATAGGA (1431-1444) and its complement TCCTATTATGTTCC (1741-1754). What function they may have, if any, is not presently clear.

Swift et al. (25) found that the 5' flanking region of each of five genes expressed in exocrine pancreas (elastase I, elastase II, chymotrypsin B, trypsin I and trypsin II) contained a conserved sequence, approximately 20 bp in length. The segment of the human glucagon gene from positions 63 to 73 differs by a single nucleotide from an 11 bp portion of the consensus of these regions, TCACCTGTXCT. This stretch, in the putative

5' untranslated region of mRNA contained in exon 1, is TCACCTGCTCT. It has not been determined whether this highly conserved sequence has a function.

A comparison of the putative 180 amino acid preproglucagons of rat (12), hamster (11), cow (10) and human shows that rat and hamster differ from human at 16 amino acids (8.9% divergence), while bovine differs from human at 13 amino acids (7.2% divergence). Most of these differences occur in the GRPP, GLP-2 and signal peptide regions. The glucagon and GLP-1 regions are identical in all four species. Although it is not known whether GLP-1 is released from preproglucagon in any tissue, the high degree of conservation of primary sequence among diverse species provides favorable evidence suggesting that the GLP-1 region has some critical function.

The nucleotide sequences of the bovine (10), hamster (11) and rat (12) cDNA clones and the equivalent portions (exons) of the human gene were compared to determine sequence divergence. In bovine, rat and hamster clones respectively, nucleotide substitutions occurred at 11.9%, 18.5% and 18.8% of the positions when compared to the human gene. The first exon had the largest divergence while the fourth, encoding GLP-1, had the least.

As previously reported (10-12), the sequences of exons 3, 4 and 5 (coding for glucagon, GLP-1 and GLP-2, respectively) are very similar to one another. It has been proposed that these exons arose through tandem duplication and then sequence divergence of a single glucagon coding unit (15,26). Thus, the glucagon gene stands as an example of the evolutionary flexibility of the genome, a demonstration of one way in which new protein coding sequences could be generated.

ACKNOWLEDGMENTS

This research was supported by grants from the National Institutes of Health (CA 20124), and The Robert A. Welch Foundation (G-267). J.W. was supported by postdoctoral fellowship from the National Institutes of Health (5 F32 AM07091). Funding for the Bionet computer resource was provided by a grant from the National Institutes of Health (1 U41 RR01685-01).

REFERENCES

1. Unger, R.H., Orci, L. (1981) New Engl. J. Med. **304**, 1518-1524.
2. Mutt, V., Jorpes, J.E., and Magnusson, S. (1970) Eur. J. Biochem. **15**, 513-519.
3. Mutt, V., and Said, S.I. (1974) Eur J. Biochem. **42**, 581-589.
4. Brown, J.C. (1971) Can. J. Biochem. **49**, 255-261.
5. Spiess, J., Rivier, J., Thorner, M. and Vale, W. (1982) Biochemistry **21**, 6037-6040.
6. Patzelt, C., Tager, H.S., Carroll, R.J., and Steiner, D.F. (1979) Nature **282**, 260-266.
7. Thim, L. and Moody, A.J. (1981) Regul. Pept. **2**, 139-150.
8. Sundby, F., Jacobsen, H. and Moody, A.J. (1976) Horm. Metab. Res.**8**, 366-371.
9. Ravazzola, M., Siperstein, A., Moody, A.J., Sundby, F., Jacobsen, H. and Orci, L. (1979) Endocrinology **105**, 499-508.
10. Lopez, L.C., Frazier, M.L., Su, C., Kumar, A., and Saunders, G.F. (1983) Proc. Natl. Acad. Sci. USA **80**, 5485-5489.

11. Bell, G.I., Santerre, R.F. and Mullenbach, G.T. (1983) Nature **302**, 716-718.
12. Heinrich, G., Gros, P., Lund, P.K., Bentley, R.C., and Habener, J.F. (1983) Endocrinology **115**, 2176-2181.
13. Lund, P.K., Goodman, R.H., Dee, P.C., and Habener, J.F. (1982) Proc. Natl. Acad. Sci. USA **79**, 345-349.
14. Lund, P.K., Goodman, R.H., Montminy, M.R., Dee, P.C., and Habener, J.F. (1983) J. Biol. Chem. **258**, 3280-3284.
15. Bell, G.I., Sanchez-Pescador, R., Laybourn, P.J., and Najarian, R.C. (1983) Nature **304**, 368-371.
16. Heinrich, G., Gros, P., and Habener, J.F. (1984) J. Biol. Chem. **259**, 14082-14087.
17. Schroeder, W.T., Lopez, L.C., Harper, M.E., and Saunders, G.F. (1984) Cytogenet. Cell Genet. **38**, 76-79.
18. Lawn, R.M., Fritsch, E.F., Parker, R.C., Blake, G., and Maniatis, T. (1978) Cell **15**, 1157-1174.
19. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) Molecular Cloning pp. 76-85. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
20. Southern, E. (1975) J. Mol. Biol. **98**, 503-517.
21. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA **74**, 5463-5467.
22. Messing, J. (1983) In Wu, R., Grossman, L., and Moldave, K. (eds.) Methods in Enzymology, Academic Press, New York, vol. **101**, pp. 20-78.
23. Thomas, P.S. (1980) Proc. Natl. Acad. Sci. USA **77**, 5201-5205.
24. Graves, B.J., Johnson, P.F., and McKnight, S.L. (1986) Cell **44**, 565-576.
25. Swift, G.H., Craik, C.S., Stary, S.J., Quinto, C., Lahaie, R.G., Rutter, W.J., and MacDonald, R.J. (1984) J. Biol. Chem. **259**, 14271-14278.
26. Lopez, L.C., Li, W., Frazier, M.L., Luo, C. and Saunders, G.F. (1984) Mol. Biol. Evol. **1**, 335-344.