Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 114–118, January 1989
Biochemistry

Structure of the human insulin receptor gene and characterization of its promoter

(alternative splicing/tyrosine kinase/*Alu* sequence)

SUSUMU SEINO*†, MITSUKO SEINO†, SHIGEO NISHI*†, AND GRAEME I. BELL*†‡

*Howard Hughes Medical Institute and Departments of †Biochemistry and Molecular Biology and of ‡Medicine, The University of Chicago, Chicago, IL 60637

Communicated by Rachmiel Levine, October 6, 1988 (received for review July 11, 1988)

ABSTRACT **The human insulin receptor gene, *INSR*, and its promoter region have been isolated and characterized. The gene spans >120 kilobase pairs (kbp) and has 22 exons. All introns interrupt protein coding regions of the gene. The 11 exons encoding the α subunit of the receptor are dispersed over >90 kbp, whereas the 11 exons encoding the β subunit are located together in a region of ≈30 kbp. Three transcriptional initiation sites have been identified and are located 276, 282, and 283 bp upstream of the translation initiation site. In addition, a 247-bp fragment from the promoter region possessing 62.6% of the maximal promoter activity has been identified. This promoter-active fragment lacks a TATA-like sequence but has two possible binding regions for the transcriptional factor Sp1. Comparison of the exon structure of the tyrosine kinase domain of the *INSR* with the corresponding regions of the human *SRC*, *ROS*, and *ERBB2* (*NGL*) protooncogenes indicates that the exon–intron organization of this region has not been well conserved.**

Insulin initiates a variety of metabolic effects upon binding to a specific receptor on the cell surface (1). The isolation and characterization of cDNA clones encoding the human insulin receptor indicate that the α and β subunits of the insulin receptor are derived by proteolytic processing of a common 1382 amino acid preproreceptor (2, 3). The 731-amino acid α subunit (M_r, 135,000) is external to the plasma membrane and contains the insulin-binding region. It is linked by interchain disulfide bonds to the 620-amino acid β subunit (M_r, 95,000), which includes a 194-amino acid extracellular domain, a 23-amino acid membrane-spanning segment and a 403-amino acid cytoplasmic segment that has intrinsic tyrosine kinase activity. Both the α subunit and the extracellular region of the β subunit are glycosylated.

Recent studies have indicated that the synthesis of an abnormal insulin receptor can contribute to the development of non-insulin-dependent diabetes mellitus (4, 5). As a first step in characterizing potentially abnormal genes relevant to diabetes, we have determined the exon–intron organization of the human insulin receptor gene, *INSR*, and characterized the 5′ flanking promoter sequences that regulate its expression.

MATERIALS AND METHODS

General Methods. Standard procedures were carried out as described by Maniatis *et al.* (6). Probes were labeled by nick-translation. DNA sequencing was done by the dideoxynucleotide chain-termination procedure (7) after subcloning appropriate DNA fragments into M13mp18 or M13mp19. The universal primer as well as sequence-specific oligonucleotides were used as primers. Both strands were sequenced.

Isolation of *INSR*. Segments of the gene were isolated from the partial *Hae* III/*Alu* I fetal liver library in phage λCh4A of Lawn *et al.* (8) by hybridization with fragments of the human insulin receptor cDNA (9). DNA fragments containing exons 3, 10, and 11 could not be isolated from this library. Genomic DNA blotting studies indicated that exon 3 was contained within a 7-kilobase pair (kbp) *Bam*HI fragment. *Bam*HI fragments of this size were isolated after electrophoresis in a 1% low-melting-point agarose gel and cloned in λEMBL3 and phage containing exon 3 identified by hybridization. Blotting studies indicated that exon 10 was contained within a 6.3-kbp *Eco*RI fragment. Despite repeated attempts, this *Eco*RI fragment could not be cloned. A 2.9-kbp *Xho* I/*Eco*RI subfragment was readily cloned and contained exon 11 and all but 13 bp of exon 10. The fortuitous identification, during a population study (K. Xiang, N. J. Cox, N. Sanz, P. Huang, J. H. Karam, and G.I.B., unpublished data), of a subject with a rare *Eco*RI restriction fragment length polymorphism facilitated the isolation of the remainder of exon 10. Restriction mapping suggested that there was a deletion of 1.2 kbp in the region of exon 10 in one of the two insulin receptor alleles of this individual. *Eco*RI fragments of 5–7 kbp were isolated by electrophoresis in a 1% low-melting-point agarose gel and cloned in λgt10. Four phage having *Eco*RI inserts of 5.1 kbp and containing exons 10 and 11 were isolated; the 1.2-kbp deletion in this fragment is upstream of exon 10. As no phage containing a 6.3-kbp *Eco*RI fragment from the other allele was isolated, we believe that there is a DNA sequence upstream of exon 10 that prevents the cloning of this region from normal chromosomes. The sequences of the exons and adjacent introns were determined. The positions of the exon–intron junctions were assigned by using the "GT/AG" rule (10).

Primer Extension. The start of transcription was determined by primer extension using a modification of the procedure described by Gil *et al.* (11). About 10 μg of total RNA from term placenta, adult liver, and IM-9 and Hep G2 cells and 1 pmol of the ^{32}P-labeled oligonucleotide 3′-GAGCCTCGTACTGGGGGCGCCCG-5′ complementary to nucleotides −47 to −69 of the mRNA were hybridized and extended with reverse transcriptase. The primer-extended products were separated on a 5% polyacrylamide/8 M urea gel.

Identification of the Promoter Region Necessary for Transcription of *INSR*. Various fragments of the putative promoter region were isolated and inserted into the *Bgl* II site of the bacterial chloramphenicol acetyltransferase (CAT) reporter gene plasmid pCAT3M (12). Hep G2 cells, a human hepatoblastoma-derived cell line (13), were transfected with 25 μg of plasmid DNA per 100-mm dish by using the calcium-phosphate procedure as described by Gorman (14). After 48 hr, cells were harvested, and CAT activity in cell extracts was measured (14).

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.

Abbreviation: CAT, chloramphenicol acetyltransferase.

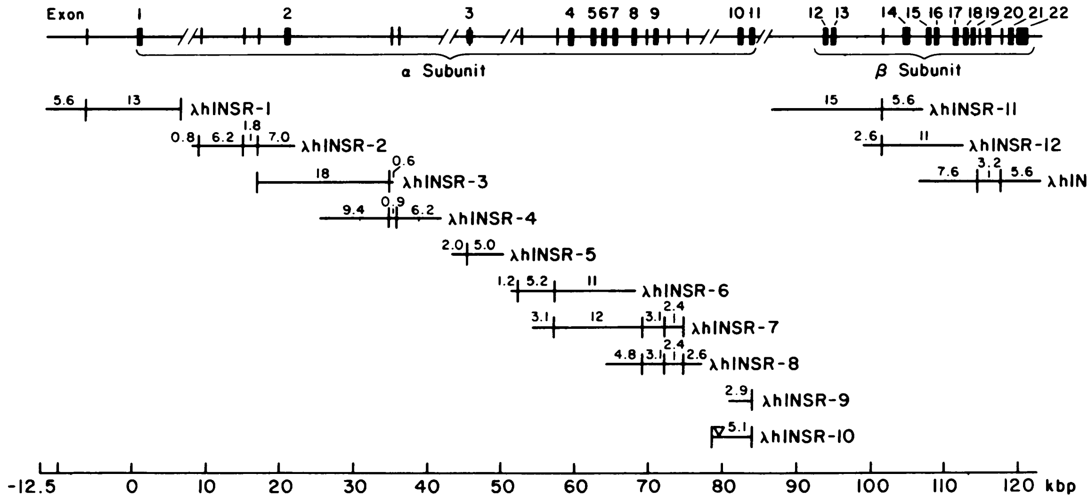

FIG. 1. Map of the human insulin receptor gene, *INSR*. The positions of the exons are indicated by the filled boxes. The transcriptional initiation site is the 0 coordinate. The name as well as the sizes and order of the *Eco*RI fragments in each clone are noted. The vertical lines represent natural *Eco*RI sites. The *Eco*RI insert in clone λhINSR-10 has a 1.2-kbp deletion upstream of exon 10.

RESULTS

Exon–Intron Organization of Human *INSR*. *INSR* was isolated as a series of overlapping DNA fragments that span >130 kbp (Fig. 1). As parts of five introns were not completely cloned, the gene must be larger than indicated in Fig. 1. The coding region of the gene comprises 22 exons (Fig. 1 and Table 1). All of the introns interrupt protein coding regions of the gene. The overall distribution of the exons is rather striking. Exons 1–3 are distributed over a region of >50 kbp. By contrast, exons 4–9 and 12–22 are located together in regions of 15 and 20 kbp, respectively, separated by a region of >25 kbp containing exon 10 and the 36-bp miniexon 11. As a consequence of this pattern of exon–intron distribution, the 11 exons encoding the α subunit of the receptor are dispersed over >90 kbp, whereas the 11 exons encoding the β subunit are located together in a region of ≈30 kbp.

Alternative Splicing Generates Different Insulin Receptors. The insulin receptor cDNA sequences first reported by Ullrich *et al.* (2) and Ebina *et al.* (3) predicted the sequences of insulin receptor precursors of 1370 and 1382 amino acids, respectively. This difference in size is due to the absence/presence of a 36-bp segment in the cDNA sequence following the codon for amino acid 717. The gene sequence indicates that this size difference results from alternative splicing of the region encoded by exon 11. Interestingly, both forms of the insulin receptor have been expressed *in vivo* (9, 15, 16) and are functional; however, the properties of the two types of proteins have not been compared rigorously.

Resolution of Nucleotide and Amino Acid Sequence Differences in the Receptor. Comparison of the *INSR* nucleotide sequence and predicted amino acid sequence of its product with reported cDNA sequences (and deduced amino acids) resolves several of the reported differences (Table 2). The

Table 1. Exon–intron organization of *INSR*

Exon	Exon size,	Sequence at exon–intron junction		Intron size,	Amino acid
no.	bp	5′ splice donor	3′ splice acceptor	kbp	interrupted
1	376/382/383	GAG G gtgagtctgg	tctcttgtag TG TGT	>25	Val-7
2	552	AAA G gtacgccggg	ctctctccag TT TGC	>25	Val-191
3	322	AGC AA gtgagttctg	ccgtccttag C TTG	>15	Asn-298
4	149	GGC A gtgagtgtct	tttgccttag AC AAT	3	Asn-348
5	145	ATT GG gtacgtgggc	ttttccatag G AAC	1	Gly-396
6	215	TCC T gtaagtcact	acttttccag GT GAA	1	Cys-468
7	127	GAG GC gtaagtagaa	gtcttgaaag C CCT	3	Ala-510
8	251	ACC A gtgagtgtgt	tccctggcag AC CCC	3	Asn-594
9	168	AAA G gtgagtgcag	ctttctccag GG CTG	>11	Gly-650
10	202	CCC AG gtcaggactt	gtttccacag A AAA	2	Arg-717
11	36	CCT AG gtatgactca	gtcgttccag G CCA	>8	Arg-729
12	275	GAA G gtagggctgc	ctccttacag CC AAG	1	Ala-821
13	140	GAT GAG gtaaggccct	ttcctcccag GAG CTG	6	Glu-867
14	160	TAT T gtaagtctcc	tctatttcag TA GAC	3	Leu-921
15	103	AAG AG gtgagttcag	cctcctccag G CAG	2	Arg-955
16	68	GAT G gtgagtacca	taagaagtag TG TTT	1	Val-978
17	245	CAC GTG gtgagtccag	tgctctgcag GTG CGC	2	Val-1059
18	111	GCT GAG gtaagctgct	tctgttttag AAT AAT	0.5	Glu-1096
19	160	GGA G gttcgtctgg	gtgttgtcag AC TTT	0.5	Asp-1150
20	130	ATG TG gtgagttgtg	tcatcggcag G TCC	1	Trp-1193
21	135	AGA GT gtaagtgtag	tgccccgcag C ACT	2	Val-1238
22	>900				

The positions at which introns interrupt the mRNA and protein sequence are indicated. Exon sequences are in capital letters; intron sequences are in lowercase. The estimated sizes of the introns are noted.

Table 2. Comparison of the human *INSR* gene, cDNA, and protein sequences

	cDNA		
Gene	Ullrich *et al.* (2)	Ebina *et al.* (3)	Whittaker *et al.* (9)
Gly--20, GGA	Gly--20, GGG	Gly--20, GGG	Gly--20, GGG
Gly-31, GGA	Gly-31, GGA	Gly-31, GGA	Gly-31, GGC
Tyr-144, TAC	Tyr-144, TAC	His-144, CAC	Tyr-144, TAC
Gln-276, CAA	Gln-276, CAG	Gln-276, CAA	Gln-276, CAA
Ile-421, ATC	Ile-421, ATC	Thr-421, ACC	Thr-421, ACC
Gln-465, CAG	Gln-465, CAG	Lys-465, AAG	Lys-465, AAG
Asp-519, GAT	Asp-519, GAC	Asp-519, GAT	Asp-519, GAC
Ala-523, GCG	Ala-523, GCA	Ala-523, GCG	Ala-523, GCG
Val-873, GTC	Asp-(861), GAC	Val-873, GTC	Val-873, GTC
Ser-874, TCC	Thr-(862), ACC	Ser-874, TCC	Ser-874, TCC
Lys-1251, AAG	Asn-(1239), AAC	Lys-1251, AAG	Lys-1251, AAG

Position numbers in parentheses differ from the others because the cDNA clone sequenced by Ullrich *et al.* (2) lacks the 12 amino acids encoded by exon 11.

differences in the codons for amino acids −20, 421, 465, and 519 represent sequence polymorphisms; those in codons for amino acids 31, 144, 276, 523, 873, 874, and 1251 need to be confirmed in other clones to exclude the possibility that they represent cloning artifacts or sequencing errors.

Sequence and Characterization of the *INSR* Promoter. As first described by Araki *et al.* (17), the promoter region of *INSR* (Fig. 2) is similar to those described for genes that are constitutively expressed (so-called housekeeping genes) in that it is extremely G+C-rich and lacks a TATA sequence. The 5′ end of the human insulin receptor transcript was mapped by primer extension to three sites 276, 282, and 283 bp upstream from the translational initiation site (the transcripts initiating at −276 and −282 are much more abundant than those starting at −283; Figs. 2 and 3). Araki *et al.* (17) have suggested that transcription is initiated at a more proximal site (indicated by "#" in Fig. 2). Unfortunately, we have been unable to confirm the primer-extension results, using an RNase-protection assay, presumably because of the G+C-rich character of the *INSR* promoter. However, the promoter-mapping studies in which Hep G2 cells were transfected with plasmid constructs containing various segments of the promoter region fused to a CAT reporter gene are consistent with the primer-extension data (Fig. 4). A 274-bp *Xho* II–*Nco* I fragment that included the region from −2 to −276 bp upstream of the translational initiation site [as well as the single putative transcriptional initiation site of Araki *et al.* (17)] had only 9.2% of the activity (a value almost identical to that obtained by using pCAT3M) of phINSRP-1, which contains 1823 bp of the 5′ flanking region. By contrast, a 247-bp fragment extending from −276 to −523 and including the three transcriptional initiation sites that we mapped by primer extension had 62.6% of the maximal promoter activity; this promoter-active fragment also contains two of four putative Spl binding regions (18) present in the 5′ flanking region of the gene. Expression of *INSR* is increased by glucocorticoids, and at least some of this effect is believed to be due to increased transcription (19). Inspection of the 5′ flanking sequence (Fig. 2) did not reveal any obvious glucocorticoid-responsive elements (20). However, there is an *Alu*

```
      Alu sequence (-1823 to -1747)
-1823 AGATCTGGCCATTGCACTCCAGCCTGGGCAACAGAGAAAAACTCCATCTAAAAAAAAAAAAAAAAAAAAAAAAAAA CAGAGAGAGAGAGAGAGAGAGAGAGAAGGAAACGGAACTGGGG

-1704 GGAGGATTTGCAAAAATATGGTTAGGGATGGCACTTCAGAGATGAAGCCATCCTGGAGTGTTACGGGCAAGGGAAATGCTGGGGCAAAGCCCCAGAGGCAGGAATAGGTTTGGCCTGTT

-1585 GCATGAACAGTGGGTCCAGCTCCTAGCAAACTGTTTATTGAATGAAAGAAGAATGAATGCCTTGGGTCTAGGGTTGTGCTGGGCGCTTTCTTAAGTTTTCTTTCCCGGGTACCTCCCCA

-1466 GAACTGGCATGCAGGTATTATTAAACCCATTACACAAGTGAAACTGGCCCAGAGACAGAAAAGTCCCTGGTCCAAGACCACACAGGAGTGAGGGGTGGAGGAACCCTCCTCCCATTGAG

-1347 TTCTGGCTTTCCTATACTGAAAGCCCCTTCCTCTCCTGCAGTAAGGTAGGTGGAACCGCTGTCCCGCCTTGTTGGTGAATGTCGTTGCTAGACTTCAGACACATACAGGCTGGTCTGCT

-1228 GAAAATCAGAGATGTCCACCTGCGCCCTATTCGAGGTCTCCGGCGTCTTCTTTGGCGTCGTCTTTGCCCTTTCAGAAGCGTCTGCACATTTTTCCAGGTGTCATTTCTCCAACTTGAAC

-1109 ACAGGGAGCGCACTGGGCACGCGGGCACGTGGCTGTCCCCAGGGGCCTGGCTTGGGTCTCGCCCCTGGGCCGGGGCGCACGCGCGGGCGGGACATCTGGGGGCGCCCACGCGCTCTGGG

 -990 ACGAGTGTCGCTGGCCAGGCCCGGACTGAGGAAAGGCGAGTGAGACACTACTCGCCTGGGGTGCAAAATTTAAGGGAGTGAAAAAAAAAAAAAAAGAAAGAAACCAAAACCACCTCGAG

 -871 TCACCAAAATAAACATTTTAATGCAGTATTTTTTAAAAAATCAACAGGAATCCTCCAAAGCCCACTATGAACAAAATAGCAAAATGGTAGAGAAAGGATCTGTGCCGCTGCGTCGGGCC

 -752 TGTGGGGCGCCTCCGGGGGTCTGAAACTGGAGGAGACTCGGGGCTGTAGGGCGCGCGGATCTGGGGCGCGCCCTCGGTCCCGGCGCGCCCAGGGCCTCCCGCGCGGGGCCCGGCACAGG

 -633 GAGGCGGGGAGGCGGGCGGGGCGGGGCGGGACCGGGCGGCACCTCCCTCCCCTGCAAGCTTTCCCTCCCTCTCCTGGGCCTCTCCCGGGCGCAGAGTCCCTTCCTAGGCCAGATCCGCG

 -514 CCGCCTTTTCCCGCGGCCCGCACGGGGCCCAGCTGACGGGCCGCGTTGTTTACGGGCCGGAGCAGCCCTCTCTCCCGCCGCCCGCCCGCCACCCGCCAGCCCAGGTGCCCGCCCGCCAG

                                                                                                                   **
 -395 TCAGCTAGTCCGTCGGTCCGCGCGTCCCTCTGTCCCGGAGCCCGCAGATCGCGACCCAGAGCGCGCGGGGCCGAGAGCCGAGAGACAGTCCCGGGCGCAGCGCGGAGCTCCGGGCCCCG

      *                                                                   #
 -276 AGATCCTGGGACGGGGCCCGGGCCGCAGCGGCCGGGGGGTCGGGGCCACCACCGCAAGGGCCTCCGCTCAGTATTTGTAGCTGGCGAAGCCGCGCGCGCCCTTCCCGGGGCTGCCTCTG

 -157 GGCCCTCCCCGGCAGGGGGGCTGCGGCCCGCGGGTCGCGGGCGTGGAAGAGAAGGACGCGCGGCCCCCAGCGCCTCTTGGGTGGCCGCCTCGGAGCATGACCCCCGCGGGCCAGCGCCG

                                              -27                         -20                                     -10
                                              Met Gly Thr Gly Gly Arg Arg Gly Ala Ala Ala Ala Pro Leu Leu Val Ala Val Ala Ala
  -38 CGCGCTCTGATCCGAGGAGACCCCGCGCTCCCGCAGCC  ATG GGC ACC GGG GGC CGG CGG GGA GCG GCG GCC GCG CCG CTG CTG GTG GCG GTG GCC GCG
                                             +1
                                  1                   7
      Leu Leu Leu Gly Ala Ala Gly His Leu Tyr Pro Gly Glu V(al) Intron 1
   61 CTG CTA CTG GGC GCC GCG GGC CAC CTG TAC CCC GGA GAG G gtgagtctgg
```

FIG. 2. Sequence of the 5′ flanking region of *INSR* and exon 1. Nucleotide numbering is relative to the first nucleotide of the codon for the initiating methionine. The three transcriptional initiation sites described in the text are denoted by asterisks [# indicates the site proposed by Araki *et al.* (17)]. The four regions representing potential binding sites for the transcriptional factor Spl are underlined. The 247-bp *Xho* II fragment having 62.6% of the maximal promoter activity extends from nucleotide −276 to nucleotide −523.

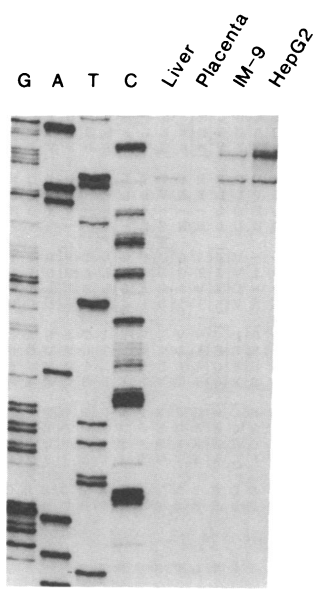

FIG. 3. Mapping the 5′ ends of human insulin receptor mRNA by primer extension. The four right-hand lanes contain the primer-extended cDNAs obtained by using RNA from the indicated tissues or cells. The sequence ladder was obtained by using the same oligonucleotide as a DNA-sequencing primer on the appropriate phage M13 template.

sequence (21) in the 5′ flanking region of *INSR*, and it is possible that this repeated sequence may represent the 5′ boundary of the promoter region.

DISCUSSION

The elucidation of the exon–intron organization of *INSR* provides an opportunity to examine the correspondence between exons and structural and functional units or modules of the insulin receptor (Fig. 5). Several of the exons encode well-defined structural units: exon 1, signal peptide; exon 2, putative ligand-binding region; exon 3, cysteine-rich region; exon 11, alternatively spliced miniexon; exon 15, transmembrane domain. The tyrosine kinase domain is encoded by five exons (exons 17–21); the region between the transmembrane and tyrosine kinase domains is encoded by a single exon (exon 16), as is the COOH-terminal hydrophilic tail of the insulin receptor (exon 22). The region encoded by exons 2–5 is also homologous to the corresponding segment of the human epidermal growth factor receptor gene, *EGFR*. As only the general organization of *EGFR* has been described (22) and the precise positions of the introns have not been determined, it is unknown if the exon structures of the homologous regions of *INSR* and *EGFR* are similar. However, intron 1 is in an equivalent position in both genes. The observed correspondence between exons and units of protein structure–function described above suggests that it is likely that exons 4–10, which encode a significant portion of the extracellular α subunit, also code for functional domains of the receptor, although their specific functions at this point are uncertain.

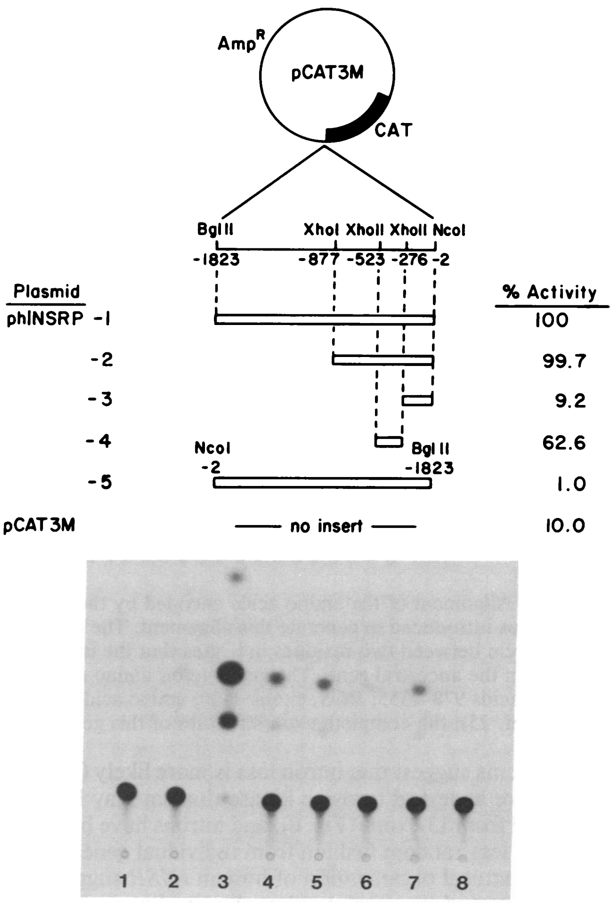

FIG. 4. Mapping promoter-active fragments from the 5′ flanking region of *INSR*. (*Upper*) The structure of each of the CAT reporter gene constructs transfected into Hep G2 cells is indicated. The relative CAT activity, represented as the mean of the results from three independent transfections, is noted. Nucleotide numbering corresponds to the sequence in Fig. 2. (*Lower*) Autoradiogram illustrating CAT activity associated with each construct (the origin is at the bottom). Lanes: 1, no DNA; 2, pCAT3M without an insert; 3, pSV2CAT, a positive control having the simian virus 40 gene early promoter/enhancer upstream of the CAT gene; 4–8, phINSRP-1 to -5, respectively.

The tyrosine kinase domain of the human insulin receptor is encoded by exons 17–21 together with a small portion of exon 22. The exon–intron organization of this domain has also been determined for the human *ROS* (23), *SRC* (24) and *ERBB2* (also called NGL for neuroglioblastoma-derived) (25) protooncogenes. Comparison of the exon structure of the tyrosine kinase domain of these four genes (Fig. 6) indicates that the exon–intron organization of this region has not been well conserved in evolution; however, there are similarities in the positions of some introns between pairs of genes. Although these data could be interpreted in terms of models involving intron loss or insertion during evolution, studies in

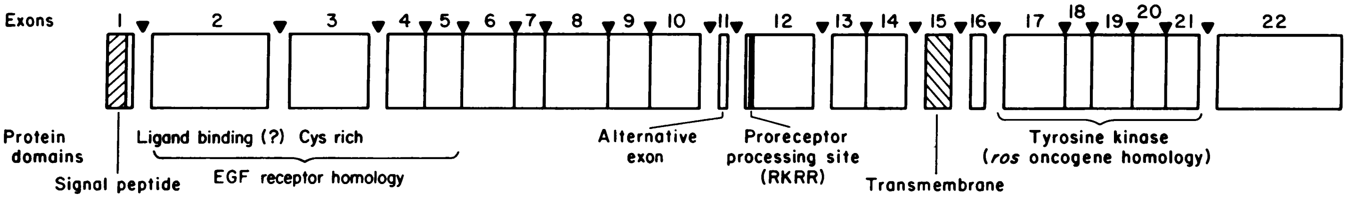

FIG. 5. Exons in *INSR* and putative protein domains of the protein. Intron positions are indicated by arrowheads. Exon numbers are shown between arrowheads. The sequence of the proreceptor processing site is shown by using the single-letter amino acid code.

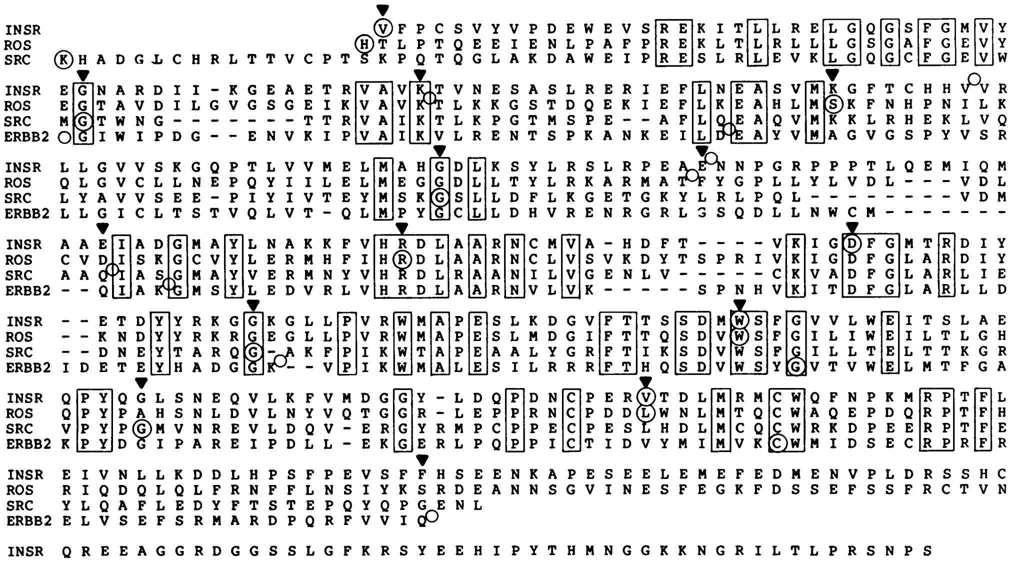
INSR V F P C S V Y V P D E W E V S R E K I T L L R E L G Q G S F G M V Y
ROS H T L P T Q E E I E N L P A F P R E K L T L R L L L G S G A F G E V Y
SRC K H A D G L C H R L T T V C P T S K P Q T Q G L A K D A W E I P R E S L R L E V K L G Q G C F G E V W
INSR E G N A R D I I - K G E A E T R V A V K T V N E S A S L R E I E F L N E A S V M K G F T C H H V V R
ROS E G T A V D I L G V G S G E I K V A V K T L K K G S T D Q E K I E F L K E A H L M S K F N H P N I L K
SRC M G T W N G - - - - - - - T T R V A I K T L K P G T M S P E - - A F L Q E A Q V M K K L R H E K L V Q
ERBB2 G I W I P D G - - E N V K I P V A I K V L R E N T S P K A N K E I L D E A Y V M A G V G S P Y V S R
INSR L L G V V S K G Q P T L V V M E L M A H G D L K S Y L R S L R P E A E N N P G R P P P T L Q E M I Q M
ROS Q L G V C L L N E P Q Y I I L E L M E G G D L L T Y L R K A R M A T F Y G P L L Y L V D L - - - V D L
SRC L Y A V V S E E - P I Y I V T E Y M S K G S L L D F L K G E T G K Y L R L P Q L - - - - - - - - V D M
ERBB2 L L G I C L T S T V Q L V T - Q L M P Y G C L L D H V R E N R G R L G S Q D L L N W C M - - - - - - -
INSR A A E I A D G M A Y L N A K K F V H R D L A A R N C M V A - H D F T - - - - V K I G D F G M T R D I Y
ROS C V D I S K G C V Y L E R M H F I H R D L A A R N C L V S V K D Y T S P R I V K I G D F G L A R D I Y
SRC A A Q I A S G M A Y V E R M N Y V H R D L R A A N I L V G E N L V - - - - - C K V A D F G L A R L I E
ERBB2 - - Q I A K G M S Y L E D V R L V H R D L A A R N V L V K - - - - - S P N H V K I T D F G L A R L L D
INSR - - E T D Y Y R K G G K G L L P V R W M A P E S L K D G V F T T S S D M W S F G V V L W E I T S L A E
ROS - - K N D Y Y R K R G E G L L P V R W M A P E S L M D G I F T T Q S D V W S F G I L I W E I L T L G H
SRC - - D N E Y T A R Q G - A K F P I K W T A P E A A L Y G R F T I K S D V W S F G I L L T E L T T K G R
ERBB2 I D E T E Y H A D G G K - - V P I K W M A L E S I L R R R F T H Q S D V W S Y G V T V W E L M T F G A
INSR Q P Y Q G L S N E Q V L K F V M D G G Y - L D Q P D N C P E R V T D L M R M C W Q F N P K M R P T F L
ROS Q P Y P A H S N L D V L N Y V Q T G G R - L E P P R N C P D D L W N L M T Q C W A Q E P D Q R P T F H
SRC V P Y P G M V N R E V L D Q V - E R G Y R M P C P P E C P E S L H D L M C Q C W R K D P E E R P T F E
ERBB2 K P Y D G I P A R E I P D L L - E K G E R L P Q P P I C T I D V Y M I M V K C W M I D S E C R P R F R
INSR E I V N L L K D D L H P S F P E V S F F H S E E N K A P E S E E L E M E F E D M E N V P L D R S S H C
ROS R I Q D Q L Q L F R N F F L N S I Y K S R D E A N N S G V I N E S F E G K F D S S E F S S F R C T V N
SRC Y L Q A F L E D Y F T S T E P Q Y Q P G E N L
ERBB2 E L V S E F S R M A R D P Q R F V V I Q
INSR Q R E E A G G R D G G S S L G F K R S Y E E H I P Y T H M N G G K K N G R I L T L P R S N P S

FIG. 6. Alignment of the amino acids encoded by the human *INSR*, *SRC*, *ROS*, and *ERBB2* genes. Identical residues are boxed. Dashes indicate gaps introduced to generate this alignment. The positions at which introns interrupt the sequences are denoted by the encircled amino acids; a circle between two residues indicates that the intron is between these two amino acids. Arrowheads indicate the proposed positions of introns in the ancestral gene. The single-letter amino acid code is used. The following regions of each gene are indicated: *INSR*, exons 17–22, amino acids 978–1355; *ROS*, exons 4–10, amino acids 137–471 (23); *SRC*, exons 7–12, amino acids 232–533 (24); and *ERBB2*, amino acids 737–990 (ref. 25); the complete exon structure of this gene has not been determined).

other systems suggest that intron loss is more likely (26). Thus, the putative ancestral tyrosine kinase domain may have been assembled from 13 exons (Fig. 6), and introns have been lost in a more or less random fashion from individual genes.

The structural organization of human *INSR* suggests that, like the structurally unrelated low density lipoprotein receptor gene (27), it is a mosaic having been constructed of exons recruited from other sources. Furthermore, as introns appear to divide *INSR* into segments that define structural and functional elements within the final encoded protein, the gene structure could provide a rational basis for creating deletions in the protein with the purpose of elucidating the function of each subregion.

We thank Drs. K. Xiang and J. H. Karam for providing DNA from the patient having the deletion upstream of exon 10; Dr. L. Laimins, Mr. D. Gius, and Ms. A. Okamoto for their advice on transfection of cultured cells and CAT assays; Mr. P. Gardner for preparing the oligonucleotides used in this study; Drs. J. Whittaker, T. Shapiro, and D. F. Steiner for comments that measurably improved the quality of this manuscript; and Ms. J. Dicig for her assistance in its preparation. This research was supported by the Howard Hughes Medical Institute.

1. Rosen, O. M. (1987) *Science* **237**, 1452–1458.
2. Ullrich, A., Bell, J. R., Chen, E. Y., Herrera, R., Petruzelli, L. M., Dull, T. J., Gray, A., Coussens, L., Liao, Y., Tsubokawa, M., Mason, A., Seeburg, P. H., Grunfeld, C., Rosen, O. M. & Ramachandran, J. (1985) *Nature (London)* **313**, 756–761.
3. Ebina, Y., Ellis, L., Jarnagin, K., Edery, M., Graf, L., Clauser, E., Ou, J., Masiarz, F., Kan, Y. W., Goldfine, I.D., Roth, R. A. & Rutter, W. J. (1985) *Cell* **40**, 747–758.
4. Yoshimasa, Y., Seino, S., Whittaker, J., Kakehi, T., Kosaki, A., Kuzuya, H., Imura, H., Bell, G. I. & Steiner, D. F. (1988) *Science* **240**, 784–787.
5. Kadowaki, T., Bevins, C. L., Cama, A., Ojamaa, K., Marcus-Samuels, B., Kadowaki, H., Beitz, L., McKeon, C. & Taylor, S. I. (1988) *Science* **240**, 787–790.
6. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
7. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143**, 161–178.
8. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. & Maniatis, T. (1978) *Cell* **15**, 1157–1174.
9. Whittaker, J., Okamoto, A. K., Thys, R., Bell, G. I., Steiner, D. F. & Hofmann, C. A. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 5237–5241.
10. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 4853–4857.
11. Gil, G., Smith, J. R., Goldstein, J. L. & Brown, M. S. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 1863–1866.
12. Laimins, L. A., Gruss, P., Pozzatti, R. & Khoury, G. (1984) *J. Virol.* **49**, 183–189.
13. Knowles, B. B., Howe, C. C. & Aden, D. P. (1980) *Science* **209**, 479–499.
14. Gorman, C. (1985) in *DNA Cloning: A Practical Approach*, ed. Glover, D. M. (IRL, Oxford), Vol. 2, pp. 143–190.
15. Ebina, Y., Edery, M., Ellis, L., Standring, D., Beaudoin, J., Roth, R. A. & Rutter, W. J. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 8014–8018.
16. Chou, C. K., Dull, T. J., Russell, D. S., Gherzi, R., Lebwohl, D., Ullrich, A. & Rosen, O. M. (1987) *J. Biol. Chem.* **262**, 1842–1847.
17. Araki, E., Shimada, F., Uzawa, H., Mori, M. & Ebina, Y. (1987) *J. Biol. Chem.* **262**, 16186–16191.
18. Dynan, W. S. & Tjian, R. (1985) *Nature (London)* **316**, 774–778.
19. McDonald, A. R., Maddux, B. A., Okabayashi, Y., Wong, K. Y., Hawley, D. M., Logsdon, C. D. & Goldfine, I. D. (1987) *Diabetes* **36**, 779–781.
20. Karin, M., Haslinger, A., Holtgreve, H., Richards, R. I., Krauter, P., Westphal, H. M. & Beato, M. (1984) *Nature (London)* **308**, 513–519.
21. Rubin, C. M., Houck, C. M., Deniniger, P. L., Friedmann, T. & Schmid, C. W. (1980) *Nature (London)* **284**, 372–374.
22. Haley, J., Whittle, N., Bennett, P., Kinchington, D., Ullrich, A. & Waterfield, M. (1987) *Oncogene Res.* **1**, 375–396.
23. Matsushime, H., Wang, L. & Shibuya, M. (1986) *Mol. Cell. Biol.* **6**, 3000–3004.
24. Anderson, S. K., Gibbs, C. P., Tanaka, A., Kung, H. & Fujita, D. J. (1985) *Mol. Cell. Biol.* **5**, 1122–1129.
25. Semba, K., Kamata, N., Toyoshima, K. & Yamamoto, T. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 6497–6501.
26. Gilbert, W., Marchionni, M. & McKnight, G. (1986) *Cell* **46**, 151–154.
27. Sudhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. (1985) *Science* **228**, 815–822.