MOLECULAR AND CELLULAR BIOLOGY, Sept. 1986, p. 3094–3108
0270-7306/86/093094-15$02.00/0

Vol. 6, No. 9

Sequence of the Notch Locus of *Drosophila melanogaster*: Relationship of the Encoded Protein to Mammalian Clotting and Growth Factors

SIMON KIDD, MARK R. KELLEY, AND MICHAEL W. YOUNG*

The Rockefeller University, New York, New York 10021

Received 12 March 1986/Accepted 16 May 1986

The Notch locus is essential for proper differentiation of the ectoderm in *Drosophila melanogaster*. Notch corresponds to a 37-kilobase transcription unit that codes for a major 10.4-kilobase polyadenylated RNA. The DNA sequence of this transcription unit is presented, except for portions of the two largest intervening sequences. DNA sequences also were obtained from three Notch cDNA clones, allowing the 5′ and 3′ ends of the gene to be mapped, and the structures and locations of nine RNA coding regions to be determined. The major Notch transcript encodes a protein of 2,703 amino acids. The protein is probably associated with cell surfaces and carries an extracellular domain composed of 36 cysteine-rich repeating units, each of about 38 amino acids. The gene appears to have evolved by repeated tandem duplications of the DNA coding for the 38-amino-acid-long protein segments, followed by insertion of intervening sequences. These repeating protein segments are quite homologous to portions of mammalian clotting factors IX and X and to the product of the *Caenorhabditis elegans* developmental gene *lin-12*. They are also similar to mammalian growth hormones, typified by epidermal growth factor.

In wild-type *Drosophila melanogaster* embryos, the nervous system develops from the part of the ectoderm called the neurogenic region. Here, all the cells appear to have the ability to become neuroblasts, but only some do so. These migrate from the neurogenic region and participate in the formation of the nervous system, leaving the remainder to become epidermal precursors (17). In neurogenic mutants this pattern of differentiation does not occur, and in the most severe phenotypes, all of the cells in the neurogenic region become neuroblasts. This results in the absence of other tissues that would have developed from the remaining epidermal precursor cells of the neurogenic region (24).

Notch is the best characterized of at least seven neurogenic loci in *Drosophila*. Null or hypomorphic mutations at Notch produce the aberrant pattern of embryonic development described above and are recessive lethals. In addition, mutations at the Notch locus affect later stages of development. Null mutations at Notch (*N* mutants), when heterozygous with a wild-type allele, cause a dominant wing-nicking phenotype, giving the locus its name. Other hypomorphic mutations at Notch are less extreme. These are recessive visibles (e.g., *spl* and *nd*) that produce eye and wing phenotypes (25). Finally, there is a class of "gain of function" dominant alleles called Abruptex (*Ax*), which produce flies that have gapped wing veins rather than nicked wing margins. Some of these alleles are also recessive lethals (55). *Ax* mutations have complex interactions both with each other and with *N* alleles. For example, some heterozygotes carrying a different *Ax* allele on each homolog show a more extreme mutant phenotype than either homozygote alone. Also, many *Ax* alleles can suppress the wing-nicking phenotype of Notch null mutations (11, 37). The Notch locus has been cloned (2, 20), and *spl*, the *N* mutations, and the *Ax* mutations appear to map, by recombination, in RNA-coding regions of the gene (20). This suggests that the diversity of function revealed by genetic analysis can be related to the activities of one or more Notch proteins.

The Notch locus produces three relatively abundant transcripts (2, 20; S. Kidd, unpublished). The most common of these is between 10 and 11 kilobases (kb) in size (2, 20) and is expressed throughout development but is most abundant in midstage embryos and in the first 24 h of pupation. The remaining two transcripts, about 9 and 10 kb, are found in adult females, unfertilized eggs, and early embryos (20; S. Kidd, unpublished). In this paper, the DNA sequence of the Notch locus is presented. Only portions of the two largest (ca. 11 and 8 kb) intervening sequences have been omitted from the analysis. The structure of the gene is such that a single large protein would be encoded by a major 10.4-kb RNA. The Notch protein is probably a transmembrane protein with an extracellular domain that is unusual, containing 36 repeating units that show extensive homology to certain mammalian clotting and growth factors.

MATERIALS AND METHODS

Nucleic acids. The Canton S genomic clones have been described (20). Oregon R genomic clones were from a cosmid library constructed in this laboratory (S. Kidd, unpublished). An *Eco*RI fragment containing the mutation fa^{swb} was isolated by cloning in lambda gt10. The pEMBL 18^+ and 19^+ vectors were constructed by inserting the polylinker of M13mp18 (32) into pEMBL 8^+ and 9^+ (7; S. Kidd, unpublished). Embryonic and pupal poly(A)$^+$ RNA was isolated as previously described (20).

DNA sequencing. The majority of the restriction fragments used for sequencing were treated with Klenow or T4 polymerase and cloned into the filled-in *Hin*dIII site of pEMBL 19^+. Plasmid DNA was prepared by the rapid boiling procedure (29) and put through an Elutip-d column (Schleicher & Schuell) before use. After cleavage with appropriate restriction enzymes and passage though an Elutip-d, DNA was treated with exonuclease III exactly as described by Henikoff (18). Samples containing 0.5 to 1 μg of DNA were

* Corresponding author.

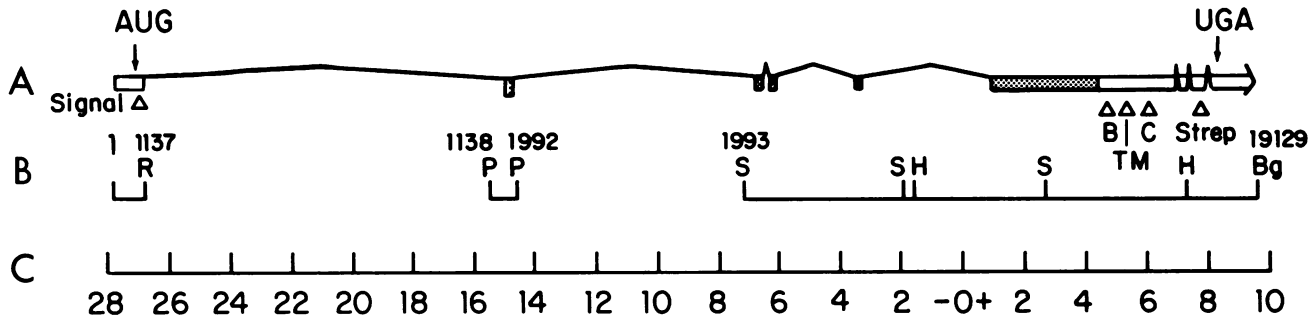

FIG. 1. Structure of the Notch transcription unit. (A) Structure of the 10.4-kb poly(A)$^+$ RNA. The position of the initiation and termination codons are indicated by AUG and UGA, respectively. The stippled parts of the exons show the positions of the EGF-like repeats. Below the map, arrows point to the location of the signal sequence (signal), transmembrane sequence (TM), and the other repetitive regions (B, C, and strep) in the Notch locus. (B) Strategy for sequencing Notch. The sequenced portions of the locus are designated by the horizontal lines. The positions of the restriction fragments cloned into pEMBL 18$^+$ and 19$^+$ from which deletions were generated are shown on the line. Only restriction sites used to produce deletions for sequencing are shown. Symbols for enzymes: R, *Eco*RI; S, *Sac*I; P, *Pst*I; H, *Hin*dIII; and Bg, *Bgl*II. The 5'-most region was sequenced left to right using deletions produced from the *Bgl*II site at −29.2 (20). (C) Coordinate scale for the Notch locus (in kilobases). The zero coordinate is at the first *Eco*RI site in Canton S proximal to the *In(1)N*^{76b8} breakpoint at Notch (20). Negative numbers go towards the telomere, positive ones towards the centromere.

removed every 32 s and mixed with an equal volume of 10 mM Tris (pH 8)–2 mM EDTA to stop the exonuclease III digestion. The volume was adjusted to 100 μl by adding a mixture giving a final concentration of 10 mM Tris, pH 8, 10 mM EDTA, 30 mM KCl, and 6 U of exonuclease VII per ml. After 1 h at 37°C, the DNA was precipitated by adding a mixture containing 1 μl of 5-mg/ml *Escherichia coli* tRNA, 20 μl of 8 M ammonium acetate, and 120 μl of isopropanol. After being frozen in a dry ice-methanol bath, the precipitate was collected by centrifugation and suspended in 10 μl of 10 mM Tris (pH 8)–0.1 mM EDTA. An equal volume of 2× T4 polymerase buffer (29) containing 1 to 3 U of T4 polymerase was added, and the samples were incubated at 37°C for 10 min. The reaction was stopped by incubation at 70°C for 10 min, and the mixture was then brought to 50 μl and final concentrations of 70 mM Tris, pH 8, 10 mM $MgCl_2$, 1 mM ATP, 15 mM dithiothreitol, and 1 U of T4 ligase. The samples were incubated overnight, and 5 to 10 μl was used to transform *E. coli* HB101 F$^+$ Kan^r (constructed by D. Ish-Horowitz).

Individual colonies were picked and spread 24 to an 8-cm plate and incubated overnight. A streak of cells was then transferred to 1.5 ml of 2× YT medium containing 0.2% glucose, 100 μg of ampicillin per ml, and 10^{10} PFU of IR1 gfip (M. Russel, unpublished) helper phage per ml and incubated with shaking at 37°C for 4 h. Single-stranded phage DNA was then extracted exactly as described previously (43) and suspended in 20 μl of 10 mM Tris (pH 8)–0.1 mM EDTA. A 5-μl sample of each phage was sequenced by using the procedure of Biggin et al. (3) except that the reactions were carried out at 42°C. Sequencing reactions were analyzed by either conventional or buffer gradient 6% acrylamide sequencing gels.

The synthetic oligonucleotide used in the primer extension experiment was synthesized on an Applied Biosystems oligonucleotide synthesizer. The primer was end-labeled with [^{32}P]ATP T4 kinase and annealed for 12 h with 20 μg of poly(A)$^+$ RNA. The nucleic acids were precipitated and divided into four samples, and the primer was extended with avian myeloblastosis virus reverse transcriptase (Bethesda Research Laboratories). The reactions were analyzed on either 6 or 10% acrylamide sequencing gels.

Computer analysis. The sequence was assembled and analyzed by using programs produced at the Biomathematics Computation Laboratory, University of California at San Francisco (30), and the DBCOMP (46), DIAGON (47), and ANALYSEQ (48) programs of Staden. The Newat (8), Protein Information Resource, and translated Genbank (provided by S. Pinsky) libraries were searched for homology to Notch by using FASTP (26). For pairwise comparisons of homologous proteins, a modification of FASTP, FASTP2 (S. Pinsky), was used. This program uses a window size of 128 rather than 32 (26) for producing an optimal alignment.

RESULTS

Fine structure of the Notch transcription unit. A physical map of the Notch locus is presented in Fig. 1. Notch corresponds to a 37-kb transcription unit that contains nine RNA coding regions. Because of the large size of the Notch locus the complete sequence of the two largest intervening sequences was not determined. The remainder was sequenced by cloning large restriction fragments (Fig. 1B) into pEMBL 18$^+$ and 19$^+$, generating deletions, and sequencing by the dideoxynucleotide procedure (see Materials and Methods). The sequence was determined on both strands and is presented in Fig. 2.

In addition, two embryonic cDNA clones homologous to RNA coding regions 1 through 6 were isolated from an Oregon R library constructed by L. Kauvar (36) and were partially sequenced. A third embryonic cDNA homologous to RNA coding regions 6 through 9 was isolated from an Oregon R library made by M. Goldschmidt-Clermont; parts of this were also sequenced. The cDNAs, together with the results of previous S1 mapping experiments (20), were used to position nine RNA coding regions that produce a major 10.4-kb transcript. Although nine coding regions were previously reported at Notch (20), there are two important differences in the Fig. 1 map. An additional 91-base-pair (bp) coding region is found at −15, and no evidence was found in the cDNA clones for the 740-bp RNA coding region previously mapped between −2.9 and −2.16. Also, no open reading frames of this size were found in the corresponding genomic DNA. Presumably this region does not encode a portion of the major Notch transcript.

The distal boundary of the Notch locus is marked by the mutation *fa*^{swb} (56). This is a small deletion which produces

FIG. 2. Sequence of the regions homologous to the 10.4-kb RNA. The sequence was determined from genomic clones previously isolated from the Canton S strain (20), with the exception of C-11148, which is from a genomic clone isolated from an Oregon R cosmid library (S. Kidd, unpublished). This residue is missing from the Canton S clone, causing sequences downstream to be out of frame when translated. The sequence was assembled and analyzed by computer as described in Materials and Methods. In the figure, exons are shown in uppercase and introns and flanking sequences in lowercase letters. Because the 3′ end of the Notch transcript was not sequenced, it was positioned by using previous S1 mapping data (20) 20 bp downstream of the poly(A) addition site at 18687. This is consistent with the location of several eucaryotic transcription termination sites (4). In numbering the DNA sequence, no allowance is made for the unsequenced portions of introns 1 and 2. In the DNA sequenced, underlined portions indicate the position of putative CAT and TATA boxes, the position of the oligonucleotide used to map the 5′ end, repeated DNA sequences upstream of the first two AUG codons in the translated part of the gene, and the direct repeats flanking introns 3 and 5. In the protein sequence, the hydrophobic signal and transmembrane sequences are shaded, the repeated EGF-like regions are in solid boxes, two other repeats (type B and C repeats; see the text) are indicated by dashed boxes, and strep is overlined. Asparagine residues of potential glycosylation sites are marked by asterisks.

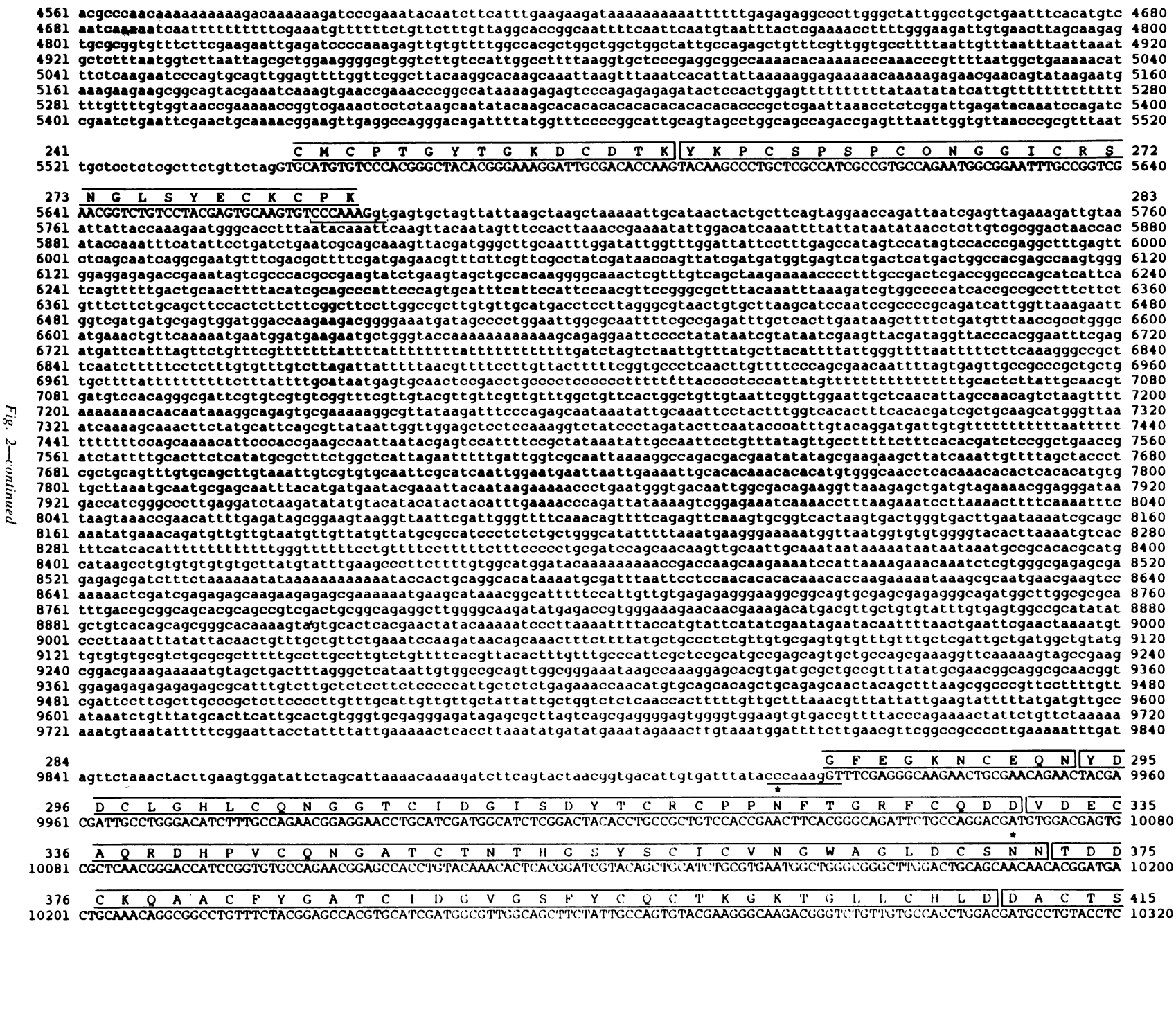

Fig. 2—continued

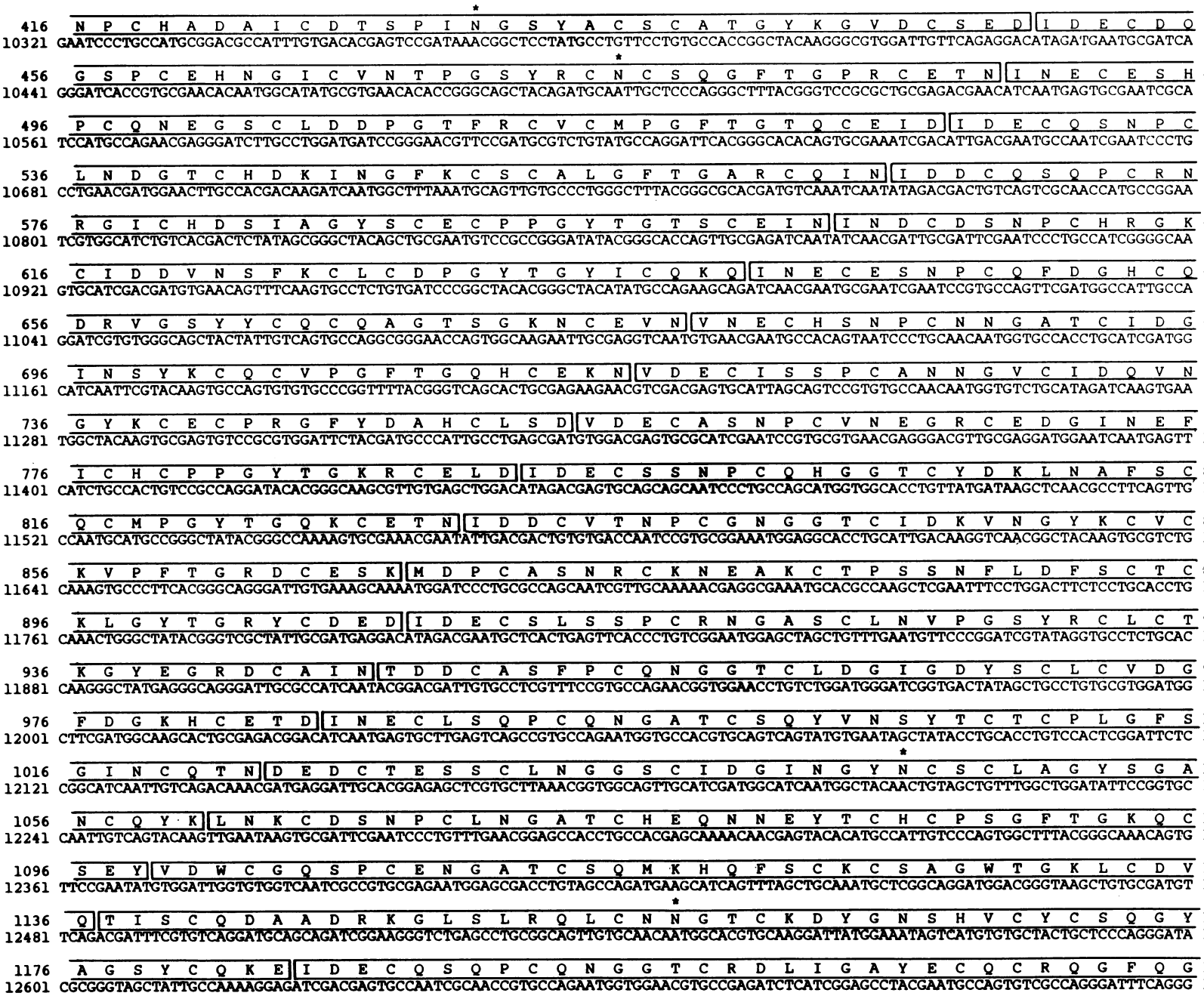

Fig. 2—continued

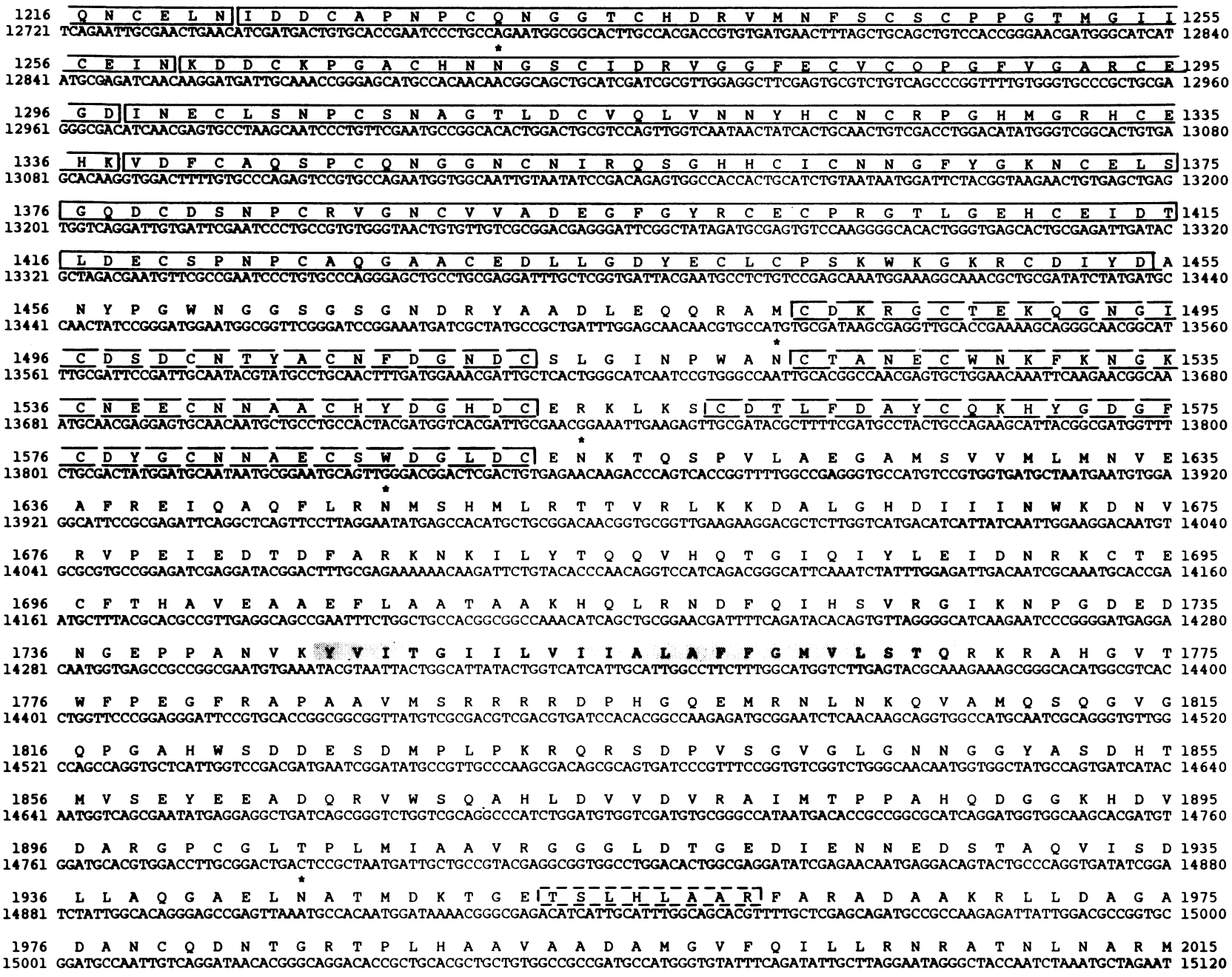

Fig. 2—continued

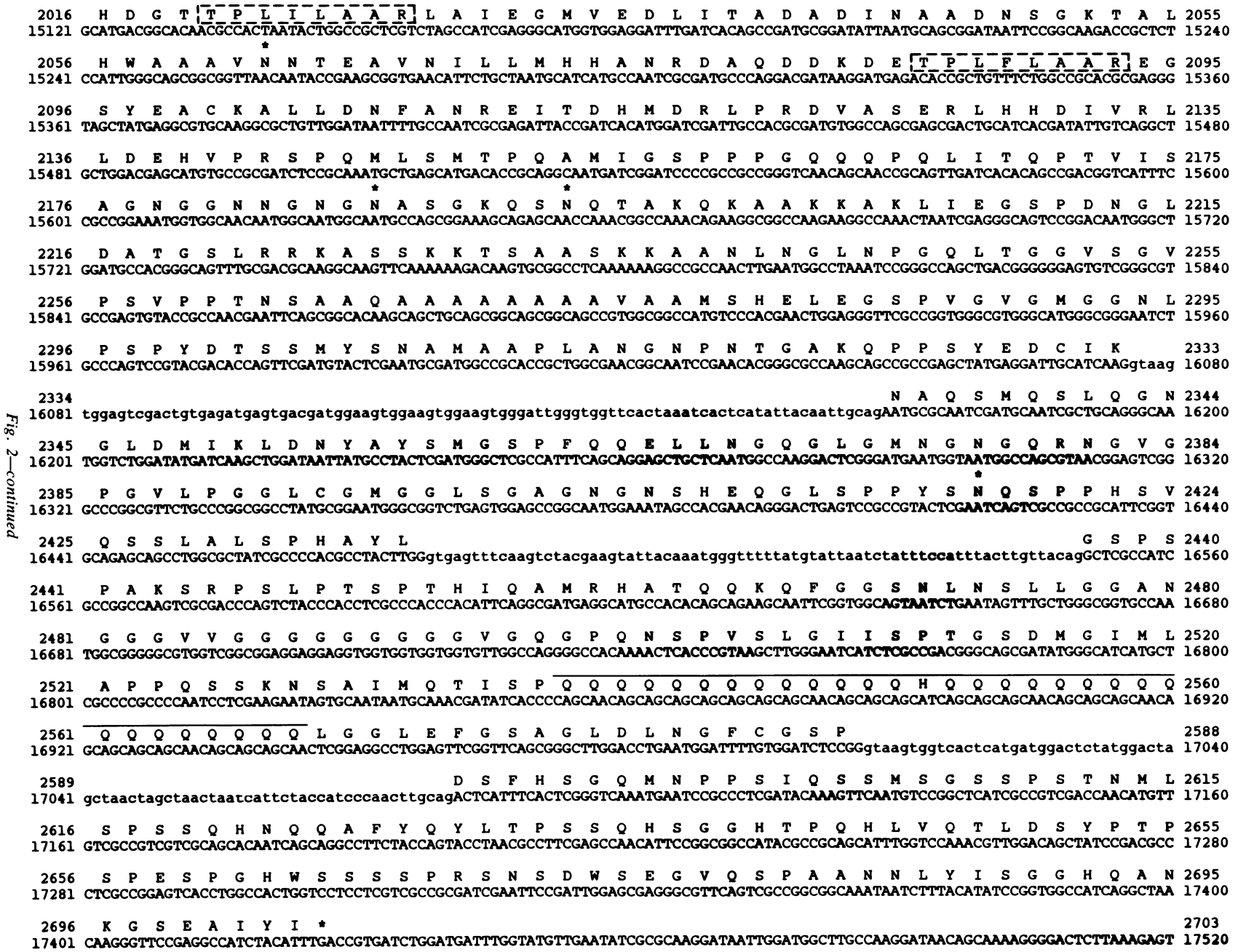

Fig. 2—continued

```
17521 CCGCGGCTAAGTTCGATCTAAAATATGCTATATATTGTAATTCTCTAATCTCTACGGATTATGTATTTTTTCCAACGACTCTTTTTTTTTTTTTTTGAGTGTGTCTATAAGTTTTAGATG 17640
17640 TTAGTGCGAAATACGCAATTCTAGCGGAAGAAGCGTCATAAATTGTAGTAACTTAAATTTTATTTAATGCTTGTTGTACAGTGCGCGCCGATCGCATATATATATATATATATAGATATT 17760
17761 GATATAGATATATGTATATCTACTCATAGATATACCAGTTAAGAGCTCGTCTTGTGGCGCATTTTATTTTATTGTCGTAACGTCATCGTAGCTTCCAGTAAAAAAAAAAAAACAAAAAACA 17880
17881 AAAGAATCAACATAAAAAGATGAGCAACTAAACTTAGCGCAAATTCTGATTGGATCGATTAAACGTTTGTGGACAATTTGTAGTTGTAAGTGACCATTATATCACTAGGCCATAAGACT 18000
18001 ACGCTAAGTTTTTTTTTATCGAATTCTACATTATTAATTATGGCCCCACATGCTATATATAAATACATAAATTTCGTAGTTCTGTAGGTTTGAAACTGAAGTGCTTATATAAACAAAATTG 18120
18121 TTAGACCGTAGACCGCATCGAACCTTTTTTTTTACACCAAGTTTTTTTTTTTTTATTATTATTATTATTATTATGACTATTATGTATTATTACATATTGACTAAGCTAAACTGTAAAAAGA 18240
18241 TTCTCATAACTGTTTGATTTAAGTTGACACATACAAAATACAAAAGGATTTGCAAAAAAAATTACACAAAAATCACCAATGGAAACGTATAAGTCTAATTGAAAGAAAGAAAGAAAAGAA 18360
18361 GAAAAAAAAGGAAAAAACCACGAAAAAGACAGTAACCAGCGAAGTTTACTATATATATAAGGACACTCGACGATCTACGATCTAAGTTGATGAAAAGATATTTTCAAAAAATATGCAACA 18480
18481 GAAAGAAAAAAAAACGCACACACTTAGAGAAAACTGCAAAATACGAATTTTCTCGTTCGATTTGCATTAATTTTAATTTGAACGGTATTGCTTTCTTTCGTCTATATTTTGTGTGTAAGTT 18600
18601 TTTAATGTGTAAATTGTAAAATTATTTGTTTTCCTAACTTAAACAGCCAGACAGCAAAACGAAACTGCAAACTTTTTGACATAAATAATAAATATTAAATAATTATTTTAAAAGCCGTGT 18720
18721 GTTTATTTCATTAATTATATACATATACTACATATAAAGGCGAATGTTTCAGCTTTGCTTTGTTGCAGCTGGAACAAAGTTTATATACCATATTTCATATTGGAACTGCGTGGAGTTAAC 18840
18841 CTTTAACCTTTAACCCTTGACACACACACACAAACCCGCAGGAAAAGAGCAGTTTTTTCGCGTGTCAAAGTGTAGATTTAATTGCAAGTTAATTATACATTAGACCTTCTTGCCGCTCTT 18960
18961 CAAGGCCACGCCCCCTCCCCCTCCTTTACCAAACGATCGCCGCCCACCACGTTGGATTTGTGTCCTTGTGTTGTGTGCGTTGTGTTTTCCGCTCTTATGTCAAAACATAACAAAAAAAAA 19080
19081 AATGTACTATGTGTTAAATTACAATATTTGGAACATTTCGACTAGATCT 19129
```

Fig. 2—continued

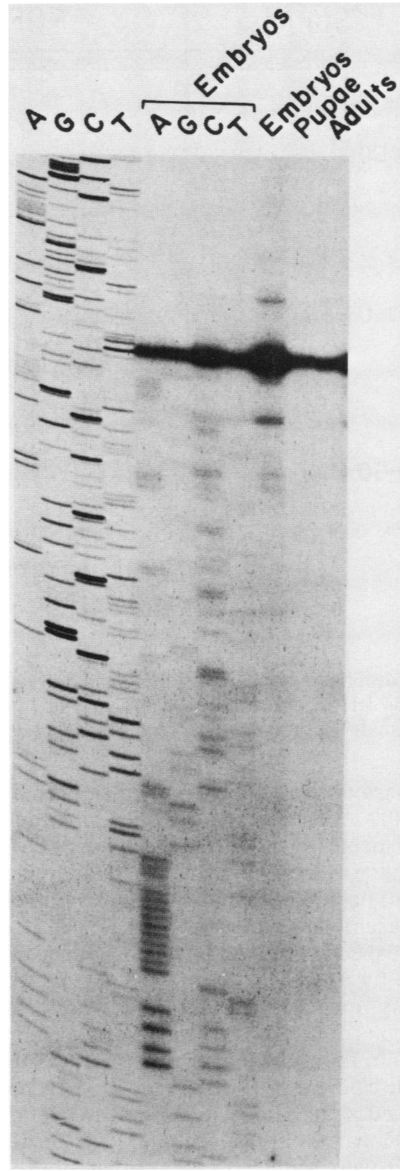

FIG. 3. Mapping the transcription start site. First four lanes: An oligonucleotide primer (see Fig. 2 for primer sequence) was annealed to cloned Notch DNA and extended according to the dideoxy sequencing method (3, 44). Last three lanes: ^{32}P-end-labeled primers were hybridized to total *Drosophila* poly(A)$^+$ RNA from the indicated stages (the adult sample was from females) and extended with reverse transcriptase (see Materials and Methods). Center four lanes: ^{32}P-end-labeled primers were hybridized to total embryo poly(A)$^+$ RNA and extended in the presence of appropriate dideoxynucleotides to generate a sequencing ladder directly from the RNA.

a rough-eye phenotype. It can be suppressed by linking fa^{swb} to more distal X-chromosomal deletions in *cis*, suggesting that the phenotype is due to a position effect (57). fa^{swb} was characterized by cloning and sequencing and corresponded to an 880-bp deletion. The position of this deletion is indicated in Fig. 2. Previous S1 mapping has placed the 5′ end of Notch just to the right of the site of the fa^{swb} deletion. A synthetic oligonucleotide primer (see Fig. 2 legend) has been extended with reverse transcriptase to map precisely the 5′ end of the transcription unit (Fig. 3). Transcription starts 67 bp from fa^{swb} at position 78 in Fig. 2. The major

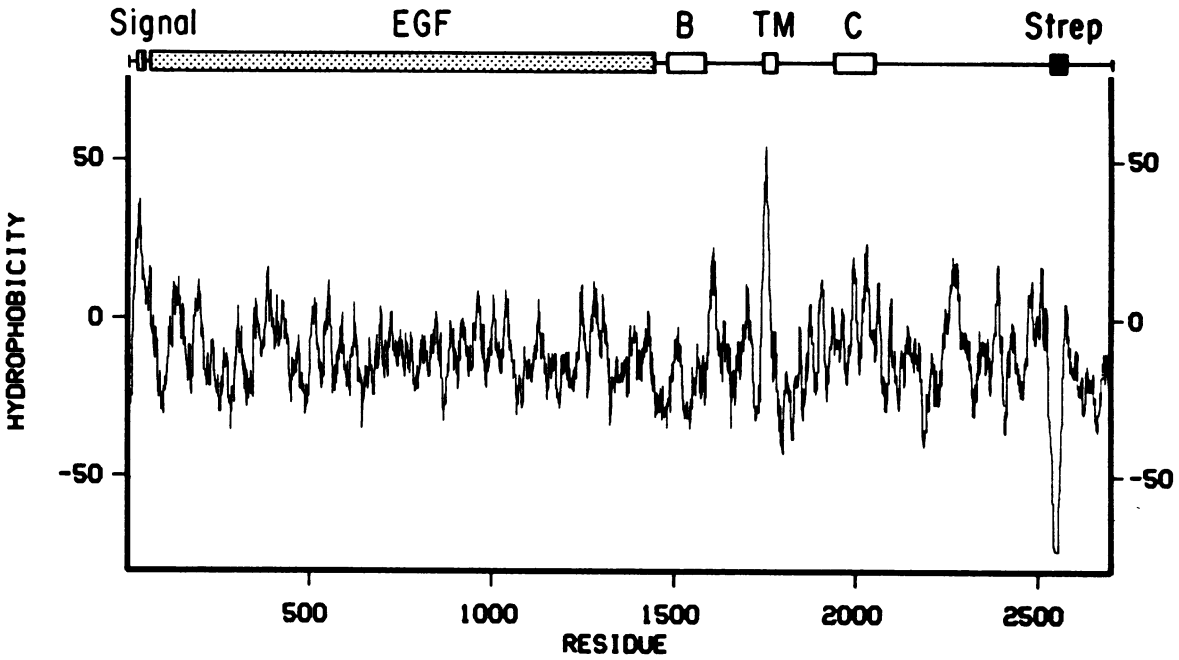

FIG. 4. Hydrophobicity profile of the Notch protein. The hydrophobicity values on the *Y* axis were calculated for 21-amino-acid windows by the method of Kyte and Doolittle (22) and plotted against the residue number of the *X* axis. Above the plot is a diagram of the Notch protein, showing its various features (see the text; see the legend to Fig. 1 for abbreviations and symbols).

10.4-kb transcript has been observed in embryos, pupae, and adults (2, 20). As shown in Fig. 3, initiation of Notch transcription occurred at the same nucleotide position at each of these developmental stages. There are sequences in this interval at positions 45 and 69 bp upstream of the 5′ end with some resemblance to consensus TATA and CAT (6) sequences, respectively. The latter sequence would be partially deleted by fa^{swb}.

The largest open reading frame in the 10.4-kb Notch transcript begins at an AUG codon 800 bp from the 5′ end; a second AUG in the same frame, 869 bp downstream, could also be used. The DNA regions surrounding both initiator codons showed some similarity to the translation initiation sequence of Kozak (21). In addition, the sequences upstream of each AUG were similar to each other; these are underlined in Fig. 2. Following the first AUG codon, an open reading frame stretches for 8,109 bp into exon 9, 1,262 bp upstream of the putative poly(A) addition site.

Organization of the protein encoded by the major Notch transcript. An open reading frame of 8,109 bp would produce a protein of 2,703 amino acids. The predicted protein was examined for hydrophobic regions by the procedure of Kyte and Doolittle (22) (Fig. 4). Two regions of high hydrophobicity were visible; the first region, between amino acid residues 23 and 44, had a structure characteristic of a signal sequence, and the second, between residues 1745 and 1767, was similar to a transmembrane protein segment (42). This suggests that Notch is a membrane-bound protein. There was a very hydrophilic region near the C terminus of the protein (residues 2538 through 2568). This corresponds to the position of the repetitive DNA sequence strep (20), also referred to as M (40) or Opa (59), which when translated consists almost entirely of glutamines.

The N-terminal 1,600 amino acids of Notch were extremely cysteine-rich, 14.6% versus 0.7% for the remainder of the protein. In the dot matrix of Fig. 5, where the Notch protein is compared with itself, the cysteine-rich region of the protein appears to be highly repetitive. The repeats could be divided into two classes, one of ca. 38 amino acids repeated 36 times, and another of of ca. 31 amino acids repeated three times. Three additional repeated regions were seen, all occurring after the transmembrane sequence. The first had three repeats, each of eight residues, between positions 1953 and 2093. The next was a 26-amino-acid sequence at position 2151 which was 50% homologous to another 26-amino-acid sequence at 2428. The final repetitive region was the polyglutamine sequence of strep. All of these repeated regions were also seen when the DNA sequence was compared with itself by using dot matrix (data not shown).

Homology of Notch protein to clotting and growth factors. Members of the first three classes of repeated sequences mentioned above are shown appropriately aligned in Fig. 6, together with consensus sequences derived from each of them. The second and third classes of repeats are referred to as type B and C repeats, respectively. The first class of repeats (36 members) was found to be homologous to an approximately 40-amino-acid, cysteine-rich sequence found in many growth and clotting factors (5, 9) and also in the product of the *C. elegans* homeotic gene *lin-12* (14). These growth and clotting factors have previously been related to each other by the homology of their cysteine-rich subsegments, and these subsegments have been referred to as epidermal growth factor (EGF)-like, indicating a limited homology with the mammalian peptide hormone EGF (9). Many of these proteins contain more than one EGF-like subsegment. For example, the EGF precursor (EGFP) has 10 (9, 13, 45), and the low-density lipoprotein receptor has three (41); many of the clotting factors contain two (34), while *lin-12* has at least 11 (14). In most of these proteins, the EGF-like sequence is known to be extracellular. In clotting factors, the sequence is part of the noncatalytic region of the protein involved in substrate recognition, and growth factors have an extracellular function in recognizing and binding to

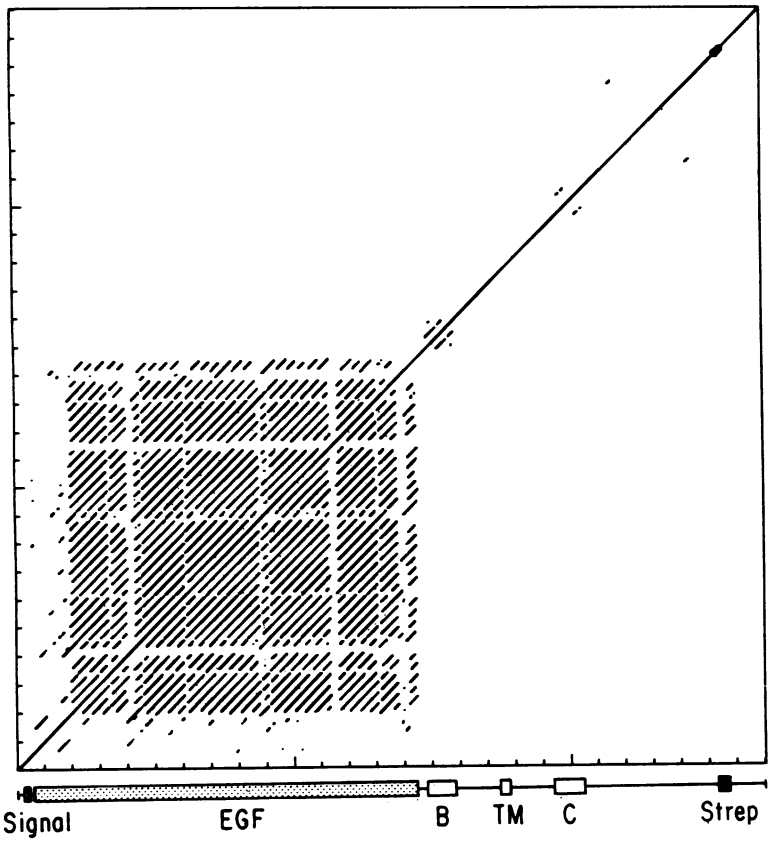

FIG. 5. Repeated sequences in the Notch protein illustrated by comparing the Notch protein with itself by using the dot matrix program (DIAGON) of Staden (47). The program has been set to scan windows of 37 amino acids and place a point when the score is greater than or equal to 430. Below this is a diagram of the Notch protein (see the legend to Fig. 1).

their receptors. These sequences are shown aligned in order of homology with the Notch consensus sequence in Fig. 7. The Notch consensus was most homologous to the first repeat of human clotting factor IX; they were the same size and had 24 of 38 residues in common. Similarly, Notch was the same size as the 10th repeat of *lin-12* and shared 18 residues. In contrast, two gaps had to be introduced into Notch to maximize its homology with growth factors; after this adjustment, it was found to share 16 residues with mouse EGF. When the remaining repeated sequences and the rest of the Notch protein were tested for similarity with other proteins, only short regions of homology were found, the significance of which is not yet clear.

Does Notch code for more than one protein? As described in the Introduction, mutations in the Notch locus affect a wide range of phenotypes. Diversity of function might be generated through developmentally regulated differential splicing of Notch transcripts, for example, by using exons containing EGF-like sequences not present in the 10.4-kb RNA. This possibility was explored by comparing the sequences from introns 1 through 5 and exons 2 through 5 against the DNA sequence of the 36 EGF-like repeats by a dot matrix. Although the 91-bp exon 2, which contains the most divergent repeat, was easily detected by this procedure, no novel homologs were found (data not shown), suggesting that if there are additional exons with EGF-like sequences, they are not in the sequenced portion of the locus.

It is clear from sequence analysis that useful transcripts could not be easily formed by alternating the splicing pattern of repeat-containing exons or any of the remaining RNA coding regions at Notch. The joining of noncontiguous RNA coding regions according to the splice junctions recognized thus far would almost always lead to frame shifts and termination within the EGF-like repeat region of the Notch protein. Coding region 7 could be joined to coding region 9 without changing the reading frame. This would eliminate the polyglutamine sequence of strep. Also, as intervening sequences 6 and 7 contained 93 and 72 bp, respectively, and neither intron contained a stop codon, either or both could be translated rather than excised to form a slightly larger Notch protein (Fig. 2). Transcripts having such structures have been searched for, but they have not been found.

In addition to the 10.4-kb transcript, two other Notch transcripts were found in RNA blotting experiments (20; S. Kidd, unpublished). Both of these may be maternal in origin, as they have been detected in unfertilized eggs and in adult females (S. Kidd, unpublished). The smaller of these, about 9 kb, failed to hybridize with Notch DNA lying to the left of coordinate 0 in Fig. 1 (data not shown), so it appears to lack exons 1 through 5. Whether this transcript is initiated upstream of exon 6 or is produced by alternate or aberrant splicing is unknown. The predicted translation product of this RNA would lack the signal sequence carried by the major Notch protein and the first six EGF-like repeats. The remaining transcript was only slightly smaller than the major Notch RNA (ca. 10.2 kb), and it appears to contain sequences corresponding to each of the RNA coding regions contributing to the major 10.4-kb transcript. The location of sequences responsible for the 0.2-kb difference has not been determined.

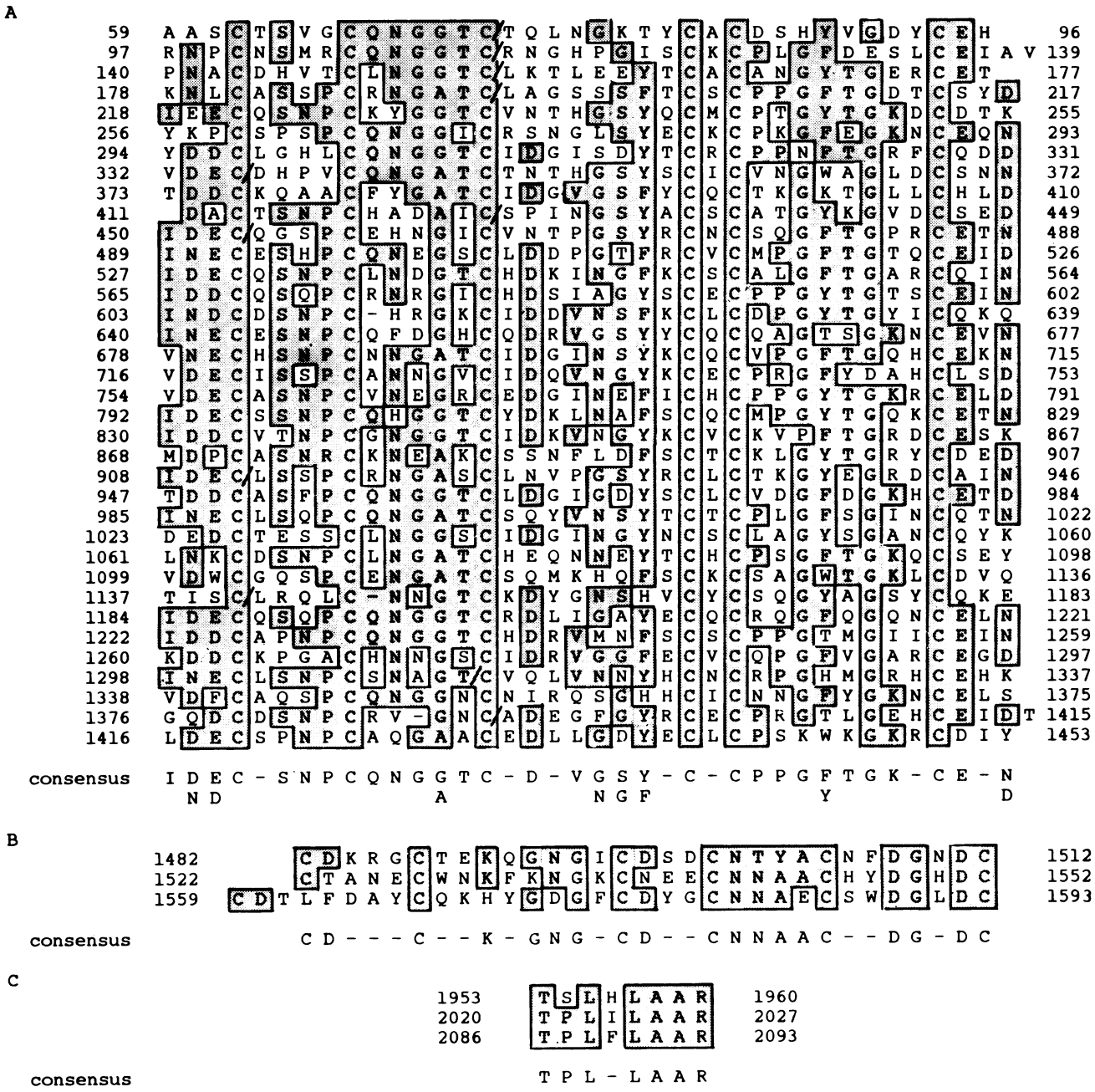

FIG. 6. Alignment of repeated sequences in Notch. The alignments were carried out with the aid of the dot matrix shown in Fig. 5, together with some run at less stringent scores. For the EGF-like repeats (A), deletions are indicated by a slash and insertions by a hyphen; this has been done to maximize homology between repeats. The boundaries are defined by comparison with factor IX (see also Fig. 7). Residues which are present in at least 25% of the EGF-like repeats are shaded. For the other cysteine-rich repeats (type B), residues present in two of the three repeats are shaded (B); likewise for the C class of repeats (C). In all cases, the consensus sequences are derived from the shaded residues; hyphens in these sequences represent any amino acid.

DISCUSSION

A total of 20 kb of the ca. 37-kb Notch locus was sequenced together with portions of three cDNA clones. Comparison of the two sequences identified RNA coding regions ranging from 91 to 6,149 bp. Notch transcription starts 67 bp downstream of the fa^{swb} deletion. Although fa^{swb} appears to delete sequences that might compose the Notch CAT box, the aberrant phenotype of fa^{swb} flies is probably not due to this loss because the phenotype can be suppressed by deleting more DNA farther upstream of Notch (57). fa^{swb} deletes the 3′ end of the transcription unit immediately upstream of the Notch locus. $Df(1)w^{67k30}$, which suppresses fa^{swb}, deletes the 5′ half of that transcription unit (S. Kidd, unpublished). Apparently, transcripts beginning outside the Notch locus and reading into it disturb Notch expression enough to give the fa^{swb} phenotype.

The predominant 10.4-kb Notch RNA is translated to produce a protein of 2,703 amino acids. While this paper was in preparation, a sequence analysis of cDNA copies of most

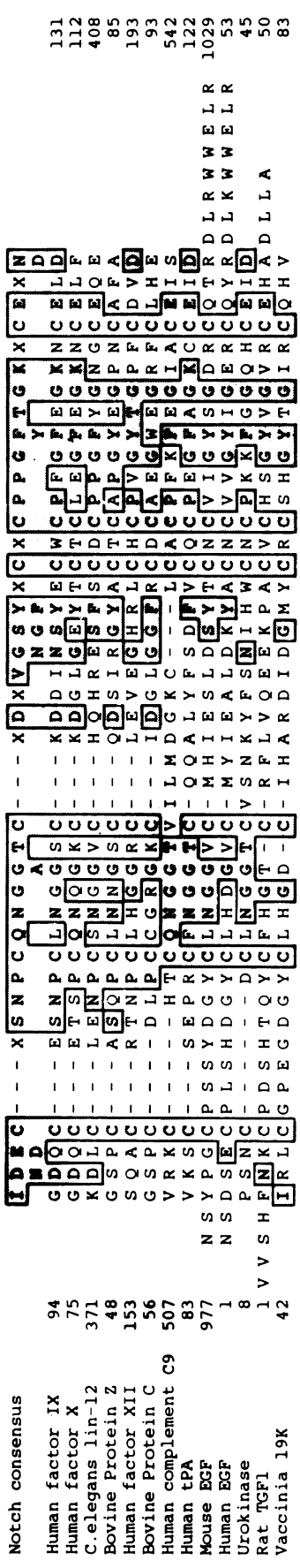

FIG. 7. Alignment of the consensus Notch EGF-like sequence with other EGF-like sequences. The EGF-like sequences were extracted from the (Protein Information Resource) data library and aligned in order of their homology to the Notch consensus sequence by using the FASTP program (26). Conserved residues are indicated by shading. The numbers on the sides refer to the position of the EGF-like sequence in the protein from which the sequence was extracted. Amino acids which are conserved between proteins are shaded. Gaps, indicated by hyphens, have been introduced in the sequences to maximize homology. Protein sequences were taken from published sequences; factor IX (1), factor X (12), *lin-12* (14), protein Z (19), factor XII (31), protein C (28), complement C9 (49), tPA (tissue plasminogen activator) (33), mouse EGF (13, 45), human EGF (15), urokinase (16), vaccinia virus 19K (53), and TGF1 (transforming growth factor 1) (23).

of this major Notch transcript was presented elsewhere (58). Most of our results, with regard to the size and structure of the predicted protein, are in agreement. However, there are a number of amino acid differences (ten). These are distributed throughout the protein and probably reflect the presence of silent-strain polymorphisms within the species.

The Notch protein encoded by the 10.4-kb RNA contains characteristic signal and transmembrane sequences and therefore is presumably a membrane-spanning protein. Its 1,745-amino-acid extracellular domain has an unusual structure, consisting of 39 cysteine-rich tandem repeats. The first 36 of these repeats, which start immediately after the signal sequence, are homologous to the EGF-like sequences found in many mammalian proteins and more recently in the product of the *lin-12* locus of *C. elegans* (14). The Notch protein repeats are extensively homologous to each other, and they are similar in size; the distance between the terminal cysteines delimiting each repeat is 32 amino acids in 63% of the repeats. Similar but lower homology can be seen at the nucleotide sequence level.

In contrast to mammalian genes, each EGF-like repeat at Notch is not encoded by a separate RNA coding region. Clusters of EGF-like repeats are separated by intervening sequences at Notch. Also in contrast to homologous mammalian genes, Notch intervening sequences are not found at conserved positions with respect to the EGF-like repeat unit. As shown in Fig. 8, all four relevant intervening sequences break EGF-like units at unique locations. Two of these (introns 3 and 5) are bounded by short, direct repeats similar to those created by the insertion of transposable elements; sequences at the 5′ and 3′ ends of these intervening sequences correspond to flanking protein-coding sequences. These direct repeats are underlined in Fig. 8. It is likely, then, that the structure at Notch evolved by tandem duplication of a single or few original repeats, followed by the insertion of introns.

By comparison of DNA and protein sequences, the least-homologous EGF-like repeats are those encoded by RNA coding regions 2, 3, and 4. These are more divergent from each other and from the remaining repeats of coding regions 5 and 6. If Notch evolved by a series of tandem duplications and insertions, then the more recently duplicated repeats may be those most homologous to each other in exons 5 and 6. Once physically separated by intervening sequences, repeats may be less prone to conservation by gene conversion or recombination or both (10). The evolution of proteins so far recognized as EGF-like in mammals probably occurred differently. Since mammalian EGF-like sequences are encoded by separate RNA coding regions (1, 33, 49–51), it has been proposed that there has been an exchange by exon shuffling between EGF sequences in different proteins (34).

The Notch consensus repeat sequence is the same size as and shows the most extensive homology to the first repeat of factor IX, one of the proteases involved in blood coagulation. This repeat is as homologous to the Notch consensus as any individual Notch repeat. Repeats with extensive homology to Notch also occur in the *lin-12* protein of *C. elegans*. A tandem array of EGF-like sequences is found in *lin-12*, and although the repeat length is more variable than that in the Notch protein, two of the repeats are similar in size and sequence to the Notch consensus, one of which is shown in Fig. 7. While Fig. 7 shows the relationship of Notch to other EGF-like proteins, the dendrogram of Fig. 9 allows several of these proteins to be compared inter se. The lengths of the lines separating the proteins in Fig. 9 are directly propor-

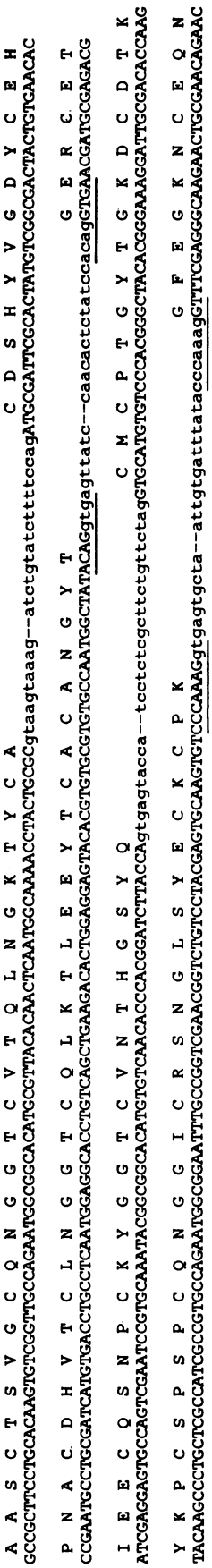

FIG. 8. Comparison of intervening sequence locations in the EGF-like domain of Notch. EGF-like sequences bordering intervening sequences 2 through 5 (top to bottom in figure) are aligned to show maximum homology of cysteine residues. The 5′ and 3′ borders of intervening sequences are shown with lowercase letters. Sequence duplications that may have resulted from insertion of intervening sequences 3 and 5 are underlined.

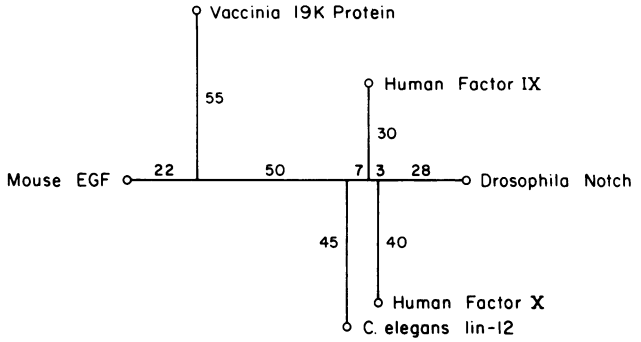

FIG. 9. Divergence of EGF-like proteins. The relationships shown are based on optimized PAM scores, accepted point mutations per 100 amino acid residues (26), generated by pairwise comparisons of the indicated proteins with FASTP2 (see Materials and Methods). Line lengths reflect PAM scores and are proportional to the degree of amino acid sequence divergence.

tional to the degree of amino acid sequence divergence. From the dendrogram it can be seen that factor IX, factor X, *lin-12*, and Notch are all about equally diverged from each other. Mouse EGF and the vaccinia virus 19-kilodalton protein are similar to each other but distantly related to Notch, *lin-12*, and the clotting factors. Because of their high level of sequence homology, common size, and wide species distribution, the EGF-like sequences in factors IX and X, Notch, and *lin-12* probably reflect the structure of the ancestral EGF-like protein. The larger sequence elements associated with mammalian growth factors, including EGF, would then have arisen through additions to this ancestral cysteine-rich sequence.

The remaining cysteine-rich sequences (type B) and the intracellular type C repeats at Notch are not homologous to any other known proteins. The strep sequence near the C terminus has already been identified in several other *Drosophila* genes (20, 36, 40, 59).

In overall structure, the major Notch protein is most homologous to the EGF precursor protein. EGFP has a signal sequence and a potential membrane-spanning domain and so is thought to be a cell-surface protein (9). The extracellular part of this protein has 10 EGF-like sequences. The C-terminal repeat in the extracellular domain is cleaved out of the precursor to produce the mature 53-amino-acid EGF molecule. There is no evidence that the other repeats are cleaved and released from the precursor. The EGFP has been identified in cells in which there is no detectable EGF (39). This has raised speculation that the EGFP may also function as a receptor molecule, particularly if it is inserted in the cell membrane. It would have a structure similar to receptors for EGF, insulin, and low-density lipoproteins (35); these all have cysteine-rich repeats in their ligand-binding extracellular domains.

Whether the Notch protein is cleaved to form peptide hormones comparable to EGF is unknown. However, if it is, then receptors for these would probably correspond to one or more of the other *Drosophila* neurogenic loci. An EGF-binding protein (52) has been identified in *Drosophila*, and a gene with homology to *erbB*, an oncogenic homolog to EGF receptor, has been cloned (27, 54). No mutations have been isolated at the cloned *erbB*-like locus, but it is not located near a chromosomal position known to be associated with a neurogenic gene. There are other restriction fragments in the *Drosophila* genome with homology to *erbB* (54), so it is still

possible that one of these corresponds to a neurogenic locus. It should be pointed out that, from genetic analysis, it is not likely that Notch produces a diffusible protein such as EGF, because Notch mutations, at least in the cuticle, are cell-autonomous; they apparently only affect the cell in which the mutation is expressed (38).

Given the cell-autonomous action of Notch, it seems most likely that Notch functions as a cell surface protein. Notch could then act as a surface-bound developmental signal or as a receptor. As a receptor, the cysteine-rich repeats would form an ordered structure due to disulfide bridges within each repeat, and by analogy with homologous sequences in mammalian proteins, the repeats may confer specific binding properties on the Notch product, so that it might interact with diffusible proteins or peptides. With regard to a possible role as a surface-bound developmental signal, the Notch protein might function in cell recognition through interactions with the surface proteins of other cells. Expression of each of the seven known *Drosophila* neurogenic loci is required to limit the number of cells committed to neuroblast formation in the embryonic neurogenic region (24). Perhaps the products of Notch and the other neurogenic loci allow groups of cells to interact through their surfaces and subsequently establish an identity that includes them in or excludes them from participation in either the neuroblast or epidermal cell lineage. This choice of one of two possible lineages is similar to the binary developmental decision governed by the *C. elegans lin-12* gene. At *lin-12* these decisions have been shown to be dependent on cell-cell interactions (14). Given the structural homology of Notch and *lin-12*, the two proteins could function in a similar manner. In any event, any protein(s) with which the Notch product interacts might again be expected to be the product of another neurogenic locus. One of the known neurogenic loci, Enhancer of split, has already been shown to interact with Notch and Delta, yet another neurogenic locus (24, 55).

A means for differentiating among some of these models may come from an analysis of hypermorphic Abruptex mutations at Notch. Comparisons of the physical and recombination maps of Notch suggest that these are apparently point mutations that lie in exon 6 (20). If the Notch protein is cleaved, releasing a single EGF-like repeat, the mutations may identify that repeat and allow the increased potency of such a protein to be understood at the molecular level. If Notch is a receptor or surface-bound signal of some type, Abruptex mutations might critically change the conformation of the cysteine-rich extracellular portion of the protein, resulting in a permanent signal being transmitted to the cell producing the protein, or changing the message that a Notch protein sends. Alternatively, *Ax* mutations may affect the intracellular domain of the protein. As many receptors have catalytic activities associated with intracellular signalling, an activity of this sort might be increased in Abruptex mutations, so they could aid in identifying such a region of the protein and any affected biochemical process.

ACKNOWLEDGMENTS

We thank Kelly Seremak for technical assistance, Elizabeth Goldstein and Peter Model for synthesizing oligonucleotide primers, and Seth Pinsky for computer programs. We thank Ted Bargiello, Laurel Eckhardt, and Norton D. Zinder for comments on the manuscript.

This work was supported by grants from the National Institutes of Health and the Andre and Bella Meyer Foundation to M.W.Y. Grant PF-2546 from the American Cancer Society supported M.R.K.

LITERATURE CITED

1. **Anson, D. S., K. H. Choo, D. J. G. Rees, F. Giannelli, K. Gould, J. A. Huddleston, and G. G. Brownlee.** 1984. The gene structure of human anti-haemophilic factor IX. EMBO J. **3**:1053–1060.
2. **Artavanis-Tsakonas, S., M. A. T. Muskavitch, and B. Yedvobnick.** 1983. Molecular cloning of *Notch*, a locus affecting neurogenesis in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **80**:1977–1981.
3. **Biggin, M. D., T. J. Gibson, and G. F. Hong.** 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80**:3963–3965.
4. **Birnstiel, M. L., M. Busslinger, and K. Strub.** 1985. Transcription termination and 3′ processing: the end is in site! Cell **41**:349–359.
5. **Blomquist, M. C., L. T. Hunt, and W. C. Barker.** 1984. Vaccinia virus 19-kilodalton protein: relationship to several mammalian proteins, including two growth factors. Proc. Natl. Acad. Sci. USA **81**:7363–7367.
6. **Breathnach, R., and P. Chambon.** 1981. Organization and expression of eukaryotic split genes coding for proteins. Annu. Rev. Biochem. **50**:349–383.
7. **Dente, L., G. Cesareni, and R. Cortese.** 1983. pEMBL, a new family of single stranded plasmids. Nucleic Acids Res. **11**:1645–1655.
8. **Doolittle, R. F.** 1981. Similar amino acid sequences: chance or common ancestry? Science **214**:149–159.
9. **Doolittle, R. F., D. F. Feng, and M. S. Johnson.** 1984. Computer-based characterization of epidermal growth factor precursor. Nature (London) **307**:558–560.
10. **Dover, G.** 1982. Molecular drive: a cohesive mode of species evolution. Nature (London) **299**:111–117.
11. **Foster, G. G.** 1975. Negative complementation at the *Notch* locus of *Drosophila melanogaster*. Genetics **81**:99–120.
12. **Fung, M. R., C. W. Hay, and R. T. A. MacGillivray.** 1985. Characterization of an almost full-length cDNA coding for human blood coagulation factor X. Proc. Natl. Acad. Sci. USA **82**:3591–3595.
13. **Gray, A., T. J. Dull, and A. Ullrich.** 1983. Nucleotide sequence of epidermal growth factor cDNA predicts a 128,000 molecular weight protein precursor. Nature (London) **303**:722–725.
14. **Greenwald, I.** 1985. lin-12, a nematode homeotic gene, is homologous to a set of mammalian proteins that includes epidermal growth factor. Cell **43**:583–590.
15. **Gregory, H., and B. M. Preston.** 1977. The primary structure of human urogastrone. Int. J. Peptide Protein Res. **9**:107–118.
16. **Gunzler, W. A., G. J. Steffens, F. Otting, S.-M. A. Kim, E. Frankus, and L. Flohe.** 1982. The primary structure of high molecular mass urokinase from human urine. The complete amino acid sequence of the A chain. Hoppe-Seyler's Z. Physiol. Chem. **363**:1155–1165.
17. **Hartenstein, V., and J. A. Campos-Ortega.** 1984. Early neurogenesis in wild type *Drosophila melanogaster*. Wilhelm Roux's Arch. Dev. Biol. **193**:308–325.
18. **Henikoff, S.** 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene **28**:351–359.
19. **Hojrup, P., M. S. Jensen, and T. E. Peterson.** 1985. Amino acid sequence of bovine protein Z; a vitamin K-dependent serine protease homolog. FEBS Lett. **184**:333–338.
20. **Kidd, S., T. J. Lockett, and M. W. Young.** 1983. The *Notch* locus of Drosophila melanogaster. Cell **34**:421–433.
21. **Kozak, M.** 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Res. **12**:857–871.
22. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.
23. **Lee, D. C., T. M. Rose, N. R. Webb, and G. J. Todaro.** 1985. Cloning and sequence analysis of a cDNA for rat transforming growth factor alpha. Nature (London) **313**:489–491.
24. **Lehmann, R., F. Jimenez, U. Dietrich, and J. A. Campos-Ortega.** 1983. On the phenotype and development of mutants of early neurogenesis in *Drosophila melanogaster*. Wilhelm Roux's

Arch. Dev. Biol. **192**:62–74.
25. **Lindsley, D. L., and E. H. Grell.** 1968. Genetic variations of *Drosophila melanogaster*. Publ. no. 627, Carnegie Institute, Washington, D.C.
26. **Lipman, D. J., and W. R. Pearson.** 1985. Rapid and sensitive protein similarity searches. Science **227**:1435–1441.
27. **Livneh, E., L. Glazer, D. Segal, J. Schlessinger, and B.-Z. Shilo.** 1985. The Drosophila EGF receptor gene homolog: conservation of both hormone binding and kinase domains. Cell **40**:599–607.
28. **Long, G. L., R. M. Belagaje, and R. T. A. MacGillivray.** 1984. Cloning and sequencing of liver cDNA coding for bovine protein C. Proc. Natl. Acad. Sci. USA **81**:5653–5656.
29. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
30. **Martinez, H. M.** 1983. An efficient method for finding repeats in molecular sequences. Nucleic Acids Res. **11**:4629–4634.
31. **McMullen, B. A., and K. Fujikawa.** 1985. Amino acid sequence of the heavy chain of human alpha factor XIIa (activated Hageman factor). J. Biol. Chem. **260**:5328–5341.
32. **Norrander, J., T. Kempe, and J. Messing.** 1983. Construction of improved M13 vectors using oligonucleotide-directed mutagenesis. Gene **26**:101–106.
33. **Ny, T., F. Elgh, and B. Lund.** 1984. The structure of the human tissue type activator gene: correlation of intron and exon structure to functional and structural domains. Proc. Natl. Acad. Sci. USA **81**:5355–5359.
34. **Patthy, L.** 1985. Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. Cell **41**:657–663.
35. **Pfeffer, S., and A. Ullrich.** 1985. Epidermal growth factor: is the precursor a receptor? Nature (London) **313**:184.
36. **Poole, S. J., L. M. Kauvar, B. Drees, and T. Kornberg.** 1985. The *engrailed* locus of Drosophila: structural analysis of an embryonic transcript. Cell **40**:37–43.
37. **Portin, P.** 1975. Allelic negative complementation at the *Abruptex* locus of *Drosophila melanogaster*. Genetics **81**:121–133.
38. **Portin, P.** 1980. On the cell lethality of *Notch* and *Abruptex* mutations of *Drosophila melanogaster*. Hereditas **92**:303–307.
39. **Rall, L. B., J. Scott, G. I. Bell, R. J. Crawford, J. D. Penschow, H. D. Niall, and J. P. Coghlan.** 1985. Mouse preproepidermal growth factor synthesis by the kidney and other tissues. Nature (London) **313**:228–231.
40. **Regulski, M., K. Harding, R. Kostriken, F. Karch, M. Levine, and W. McGinnis.** 1985. Homeo box genes of the *antennapedia* and *bithorax* complexes of Drosophila. Cell **43**:71–80.
41. **Russell, D. W., W. J. Schneider, T. Yamamoto, K. L. Luskey, M. S. Brown, and J. L. Goldstein.** 1984. Domain map of the LDL receptor: sequence homology with the epidermal growth factor precursor. Cell **37**:577–585.
42. **Sabatini, D. D., G. Kreibich, T. Morimoto, and M. Adesnik.** 1982. Mechanisms for the incorporation of proteins in membranes and organelles. J. Cell Biol. **92**:1–22.
43. **Sanger, F., A. R. Coulson, B. G. Barrell, A. J. H. Smith, and B. A. Roe.** 1980. Cloning in single stranded bacteriophage as an aid to rapid DNA sequencing. J. Mol. Biol. **143**:161–178.
44. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
45. **Scott, J., M. Urdea, M. Quiroga, R. Sanchez-Pescador, N. Fong, M. Selby, W. Rutter, and G. I. Bell.** 1983. Structure of a mouse submaxillary messenger RNA encoding epidermal growth factor and seven related proteins. Science **221**:236–240.
46. **Staden, R.** 1980. A new computer method for the storage and manipulation of DNA gel reading data. Nucleic Acids Res. **8**:3673–3694.
47. **Staden, R.** 1982. An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. Nucleic Acids Res. **10**:2951–2961.
48. **Staden, R.** 1984. Graphic methods to determine the function of nucleic acid sequences. Nucleic Acids Res. **12**:521–538.
49. **Stanley, K. K., H. P. Kocher, J. P. Luzio, P. Jackson, and J. Tschopp.** 1985. The sequence and topology of human complement, component C9. EMBO J. **4**:375–382.
50. **Sudhof, T. C., J. L. Goldstein, M. S. Brown, and D. W. Russell.** 1985. The LDL receptor gene: a mosaic of exons shared with different proteins. Science **228**:815–822.
51. **Sudhof, T. C., D. W. Russell, J. L. Goldstein, M. S. Brown, R. Sanchez-Pescador, and G. I. Bell.** 1985. Cassette of eight exons shared by genes for LDL receptor and EGF precursor. Science **228**:893–895.
52. **Thompson, K. L., S. J. Decker, and M. R. Rosner.** 1985. Identification of a novel receptor in *Drosophila* for both epidermal growth factor and insulin. Proc. Natl. Acad. Sci. USA **82**:8443–8447.
53. **Venkatesan, S., A. Gershowitz, and B. Moss.** 1982. Complete nucleotide sequences of two adjacent early vaccinia virus genes located within the inverted terminal repetition. J. Virol. **44**:637–646.
54. **Wadsworth, S. C., W. S. Vincent III, and D. Bilodeau-Wentworth.** 1985. A *Drosophila* genomic sequence with homology to human epidermal growth factor receptor. Nature (London) **314**:178–180.
55. **Welshons, W. J.** 1971. Genetic basis for two types of recessive lethality at the *Notch* locus of *Drosophila*. Genetics **68**:259–268.
56. **Welshons, W. J., and D. O. Keppy.** 1975. Intragenic deletions and salivary band relationships in *Drosophila*. Genetics **80**:143–155.
57. **Welshons, W. J., and H. J. Welshons.** 1985. Suppression of the *facet-strawberry* position effect in *Drosophila* by lesions adjacent to *Notch*. Genetics **110**:465–477.
58. **Wharton, K. A., K. M. Johansen, T. Xu, and S. Artavanis-Tsakonas.** 1985. Nucleotide sequence from the neurogenic locus *Notch* implies a gene product that shares homology with proteins containing EGF-like repeats. Cell **43**:567–581.
59. **Wharton, K. A., B. Yedvobnick, V. G. Finnerty, and S. Artavanis-Tsakonas.** 1985. *Opa*: a novel family of transcribed repeats shared by the *Notch* locus and other developmentally regulated loci in Drosophila melanogaster. Cell **40**:55–62.