Proc. Natl. Acad. Sci. USA
Vol. 75, No. 10, pp. 4853-4857, October 1978
Biochemistry

Ovalbumin gene: Evidence for a leader sequence in mRNA and DNA sequences at the exon–intron boundaries*

(split gene/mRNA splicing/eukaryotic gene structure)

R. Breathnach, C. Benoist, K. O'Hare, F. Gannon, and P. Chambon

Laboratoire de Génétique Moléculaire des Eucaryotes du Centre National de la Recherche Scientifique, Unité 44 de l'Institut National de la Santé et de la Recherche Médicale, Institut de Chimie Biologique, Faculté de Médecine, Strasbourg 67085, France

Communicated by A. Frey-Wyssling, July 31, 1978

ABSTRACT **Selected regions of cloned *Eco*RI fragments of the chicken ovalbumin gene have been sequenced. The positions where the sequences coding for ovalbumin mRNA (ov-mRNA) are interrupted in the genome have been determined, and a previously unreported interruption in the DNA sequences coding for the 5′ nontranslated region of the messenger has been discovered. Because directly repeated sequences are found at exon–intron boundaries, the nucleotide sequence alone cannot define unique excision–ligation points for the processing of a possible ov-mRNA precursor. However, the sequences in these boundary regions share common features; this leads to the proposal that there are, in fact, unique excision–ligation points common to all boundaries.**

It has been shown (1–4) that the chicken ovalbumin gene is split into seven ovalbumin messenger coding sequences (exons; see ref. 5) separated by six intervening sequences (introns; see ref. 5). The respective locations in the chicken genome of the seven exons (numbered 1–7) and of the six introns (designated by the letters B–G) are shown in Fig. 1*b*. These positions have been deduced from restriction enzyme mapping of chicken DNA using appropriate ovalbumin gene probes (1, 3, 4) and from electron microscopy of the cloned *Eco*RI DNA fragments "a," "b," "c," and "d" (Fig. 1*b*) which contain all of the ovalbumin exons and introns (2, 3). Electron microscopy did not reveal any evidence for a long (150–200 nucleotides) virus-like leader sequence (for review, see refs. 7 and 8) that could be spliced at the 5′ end of the ovalbumin mRNA (ov-mRNA) (3, 4). By comparison with viruses, we use the term "leader" to describe a nontranslated RNA sequence present at the 5′ end of a given mRNA and encoded by DNA sequences physically separated from those coding for the protein. However, the possible existence of a leader sequence shorter than 50–100 nucleotides was not excluded by our electron microscopy studies (3, 4). As discussed previously (1, 4), the split organization of the ovalbumin gene raises the possibility that the primary transcript of the gene could be longer than mature ov-mRNA and contain transcripts of both exons and introns. Maturation of ov-mRNA might then involve the looping out of intron transcripts for excision and the concomitant splicing of exon transcripts. Whatever the detailed mechanisms involved in such processing, it was generally postulated that the nature of the DNA sequences at the intron–exon junctions or in their immediate vicinity should play a role in the recognition of the intron–exon boundaries and in the excision-splicing events.

In the present paper we report the result of sequence analyses carried out both on the cloned double-stranded cDNA containing the sequences complementary to ov-mRNA (ov-ds-cDNA; see ref. 9) and on cloned cellular DNA fragments. These studies have led to the discovery of a short leader sequence at the 5′ end of ov-mRNA and have revealed some interesting features in the DNA sequences at exon–intron boundaries.

MATERIALS AND METHODS

Plasmid pCR1 ov 2.1 containing the ov-ds-cDNA insert was prepared as described (9). *Eco*RI fragments "b," "c," and "d" previously cloned in λ vectors (3) were transferred to the plasmid pBR 322. An *Eco*RI/*Hin*dIII fragment of the *Eco*RI fragment "a" containing the entirety of exon 7 (Fig. 1*b*) was also transferred to a vector derived from pBR322 by digestion with *Eco*RI and *Hin*dIII. Superhelical DNA plasmids were prepared, digested with *Eco*RI (for fragments "b," "c," and "d") or *Eco*RI plus *Hin*dIII (for the *Eco*RI/*Hin*dIII fragment) and the fragments were purified on sucrose gradients. Restriction enzyme sites in these fragments were mapped as described (3) or by the method of Smith and Birnstiel (10). For 5′-^{32}P end-labeling, fragments were digested with restriction enzymes, treated with bacterial alkaline phosphatase, and incubated with polynucleotide kinase T4 (a gift of F. Rougeon) in 50 mM Tris·HCl, pH 7.9/10 mM $MgCl_2$/10 mM 2-mercaptoethanol/7 mM K_2HPO_4 [to inhibit phosphatase (11)] containing 1 μM [γ-^{32}P]ATP (Amersham, >3000 Ci/mmol). After cleavage with a second restriction enzyme, fragments labeled at one end only were isolated by polyacrylamide gel electrophoresis. Elution from the gel was as described (12) using a DEAE-cellulose column step. In some cases, fragments labeled at both ends were first separated by polyacrylamide gel electrophoresis and then eluted and digested with a second restriction enzyme; fragments labeled at one end only were obtained by polyacrylamide gel electrophoresis. End-labeled fragments were sequenced by the chemical degradative technique of Maxam and Gilbert (13) as modified (5). Five base-specific cleavages were used (G, A > G, C + T, C, A > C). Electrophoresis was on 90-cm-long, 8 or 20% polyacrylamide gels. Autoradiography was on pre-exposed Kodak RP Royal X-Omat film with Du Pont Lightning Plus screens at −90°. Sites used for end-labeling and the extent of sequences obtained are shown in Fig. 2. Sources of restriction enzymes were as described (3, 4).

Biohazards associated with the experiments described here were examined previously by the French National Control Committee. The experiments were carried out accordingly under L3-B1 conditions in the nomenclature adopted by the French Committee (L3-B1 is considered equivalent to P3, EK1 in the National Institutes of Health nomenclature).

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.

Abbreviations: ov-mRNA, ovalbamin mRNA; ov-ds-cDNA, ovalbumin double-stranded cDNA.

[*] A preliminary account of this work was presented at the Francqui Colloquium on Differentiation (Bruxelles, Belgium, June 19–22, 1978).

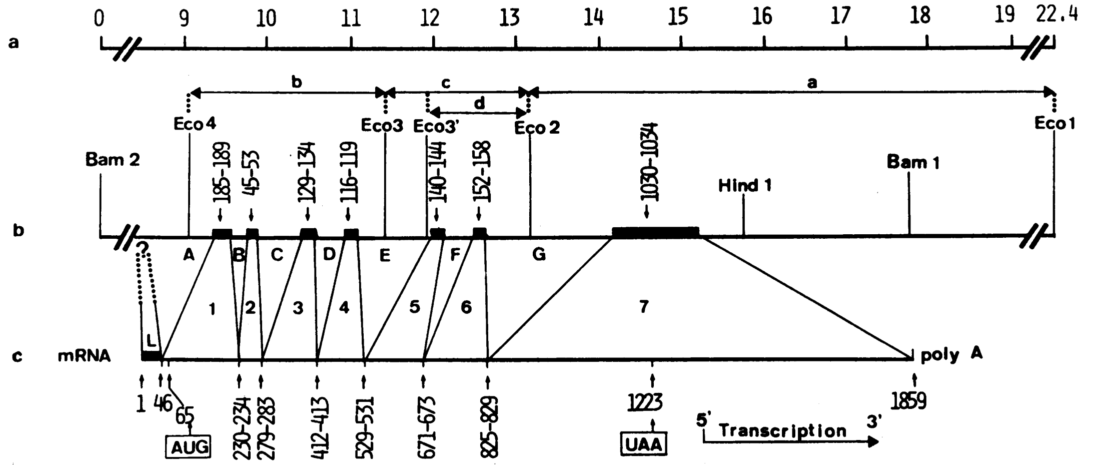

FIG. 1. General organization of the ovalbumin gene. (*a*) Scale in kilobase pairs for *b*. (*b*) Location of the exons (1–7, heavy lines) and introns (A–G) of the ovalbumin gene within the cellular 18-kilobase *Bam*HI (Bam) fragment and the *Eco*RI (Eco) fragments *a* to *d* (taken from refs. 1, 3, and 4). The *Eco*RI/*Hin*dIII fragment, which was cloned in pBR322, is defined by the *Eco*RI and *Hin*dIII sites Eco2 and Hind1, respectively. The numbers above the exons refer (in base pairs) to their maximum and minimum possible sizes (see text). (*c*) Schematic representation of ovalbumin mRNA. The total length and the positions of the AUG and UAA codons are from ref. 6. L refers to the leader sequence (see text) and the vertical numbers indicate the possible positions with respect to the mature mRNA where the exon transcripts could be ligated (see text and Fig. 4).

RESULTS

Evidence for a Leader Sequence in Ov-mRNA. It has been shown that the *Eco*RI fragment "b" of the ovalbumin gene contains the sequences coding for the 5′ region of the mRNA (1, 3). In order to determine whether this fragment contains the entirety of these 5′-terminal sequences, we compared sequences corresponding to the 5′ end of the mRNA in the cloned ds-cDNA (clone pCR1 ov 2.1, see ref. 9) and in the cloned cellular *Eco*RI fragment "b" (3). The region of the cloned ds-cDNA containing the sequences coding for the 5′ end of the mRNA was sequenced as outlined in Fig. 2*a*, extending in the 5′ direction from the *Sst* I site also present in exon 1 (Fig. 2*b*). Comparison of our sequence (shown in Fig. 3) with that of the NH_2-terminal sequence of ovalbumin (14) and with that of the ovalbumin mRNA sequence of McReynolds *et al.* (6) shows that the first 14 nucleotides of the messenger are not represented in pCR1 ov 2.1 DNA and also allows numbering of the nucleotides from the 5′ end of ov-mRNA (this numbering is used throughout). We observed three differences between the two sequenced ovalbumin mRNAs in this region: at positions 34 (C for G), 43 (A for G), and 79 (C for U), the last two destroying a *Taq* I site (14) and creating a *Hha* I site, respectively.

The cloned *Eco*RI fragment "b" was sequenced around the 5′ end of exon 1 as outlined in Fig. 2*b*. This sequence is compared in Fig. 3 with the sequence we obtained for the mRNA in this region. It is apparent that exon 1 contains sequences coding for protein starting from the initiation codon (positions 65–67) and for 19 nucleotides of the 5′ nontranslated region of the messenger. However, the first 45 nontranslated nucleotides cannot be encoded in exon 1 and, from our preliminary sequence data, are unlikely to be encoded in *Eco*RI fragment "b." Therefore, the DNA sequences coding for the 5′ nontranslated region of ov-mRNA are interrupted. This interruption is responsible for the existence of an *Xba* I site in the cloned cellular fragment "b" (Figs. 3 and 2*b*) that has no counterpart in the cloned ds-cDNA. The 45 nontranslated messenger nucleotides that are not encoded in exon 1 represent a leader sequence (as defined in the Introduction).

It is interesting to note the repeated triplet CTG (underlined in Fig. 3) close to the intron A–exon 1 boundary. Similar repeats

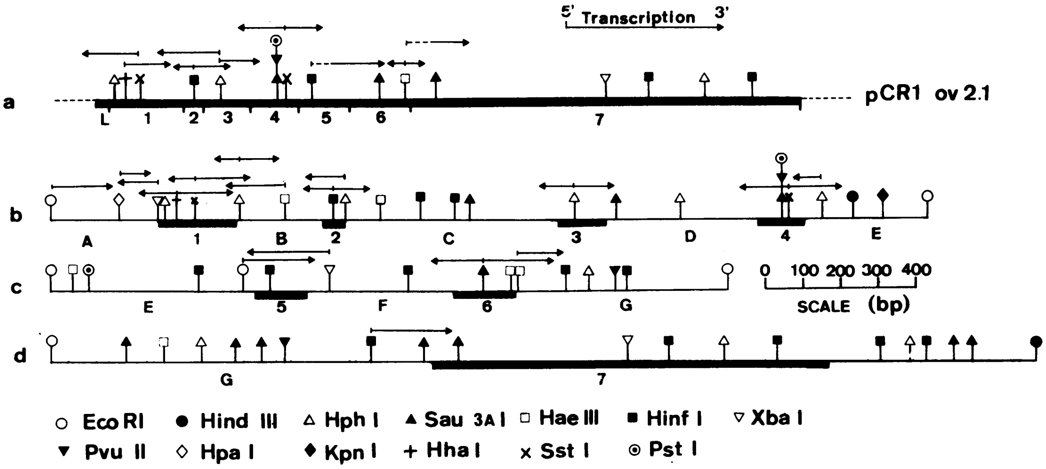

FIG. 2. Restriction enzyme maps: (*a*) cloned double-stranded ov-cDNA in pCR1 ov 2.1 (9); (*b*) *Eco*RI fragment "b" of cellular DNA; (*c*) *Eco*RI fragment "c" of cellular DNA [the inner *Eco*RI site corresponds to Eco3′ site which defines the 5′ end of *Eco*RI fragment "d" (see Fig. 1*b* and ref. 4)]; (*d*) *Eco*RI/*Hin*dIII fragment of the cellular *Eco*RI fragment "a" (see Fig. 1*b*). All of the sites for the enzymes shown are presented. The horizontal arrows indicate the direction and the extent of the sequence determinations. Numbers (1–7) and letters (L, A–G) are defined in Fig. 1.

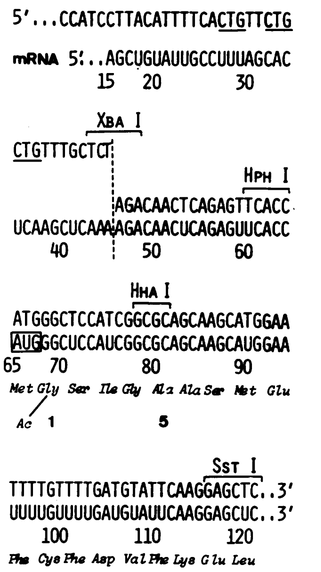
5′...CCATCCTTACATTTTCACTGTTCTG
mRNA 5′..AGCUGUAUUGCCUUUAGCAC
15
20
30
XBA I
CTGTTTGCTCT
HPH I
AGACAACTCAGAGTTCACC
UCAAGCUCAAAAGACAACUCAGAGUUCACC
40
50
60
HHA I
ATGGGCTCCATCGGCGCAGCAAGCATGGAA
AUGGGCUCCAUCGGCGCAGCAAGCAUGGAA
65
70
80
90
Met Gly Ser Ile Gly Ala Ala Ser Met Glu
Ac
1
5
SST I
TTTTGTTTTGATGTATTCAAGGAGCTC..3′
UUUUGUUUUGAUGUAUUCAAGGAGCUC..3′
100
110
120
Phe Cys Phe Asp Val Phe Lys Glu Leu
10
15

FIG. 3. Sequences showing that there is a leader at the 5′ end of ov-mRNA. The mRNA sequence was derived from the sequence in the ds-cDNA (Fig. 2*a*) coding for the 5′-terminal region of the messenger. The nucleotides are numbered with respect to the 5′ end of the mRNA by comparison with the sequence of McReynolds *et al.* (6). The DNA sequence of the noncoding strand is that of the region around the 5′ end of exon 1 (Fig. 2*b*) and contains the restriction enzyme sites *Xba* I, *Hph* I, *Hha* I, and *Sst* I (see text). The divergence of the RNA and DNA sequences upstream of position 46 is indicated by the vertical broken line. The repeated 5′-C-T-G-3′ trinucleotide is underlined (see text).

of the same triplet have been observed in a λII immunoglobulin light chain gene (5) and in the rabbit β-globin gene (A. Efstratiadis, T. Maniatis, and L. Lacy, personal communication).

DNA Sequences at the Exon–Intron Junctions. Our previous work (1–4) has shown that the interruptions in the ovalbumin gene all lie in the sequences coding for the first 900 nucleotides of ov-mRNA (see Fig. 1). We therefore sequenced more than 95% of this region in the cloned ds-cDNA as outlined in Fig. 2*a*. Our results (not shown) are in good agreement with those of McReynolds *et al.* (6). One difference was found in addition to those mentioned above: G for A at position 223. The nucleotide corresponding to position 223 is also G in the cloned cellular fragment "b."

We have extended our previous restriction enzyme maps of cloned *Eco*RI fragments "a," "b," and "c" (2–4). Our data are summarized in Fig. 2 *b–d*. Regions at exon–intron boundaries were sequenced as shown. Selected regions of the sequences obtained are shown in Fig. 4 together with the corresponding sequences of the mRNA derived from the cloned ds-cDNA. Examination of the sequences reveal the following features.

(*i*) There is no evident way to form base-paired structures that would bring into close proximity the ends of consecutive exon transcripts and to loop out intron transcripts to allow excision and splicing of a possible mRNA precursor (see *Discussion*).

(*ii*) The exon–intron boundaries cannot be uniquely defined. For example, the G at position 413 of the ov-mRNA could be encoded for either at the end of exon 3 or at the beginning of exon 4. The phenomenon becomes more marked for the remaining boundaries when two-base pair (introns E and F) or four-base pair (introns B, C, and G) indirect repeats are found (see boxed nucleotides in Fig. 4; in all cases the arrows indicate the maximum possible extent of the exons).

(*iii*) When the sequences are aligned as in Fig. 4, sequences at the intron–exon boundaries seem to fall into three types. Type 1 comprises the boundaries of introns C, F, and G and is characterized by the sequences 5′-A-G-G-T-A-3′ at the 5′ end of the introns and 5′-C-A-G-3′ at the 3′ end of the introns. Type 2 comprises the boundaries of introns D and E and is characterized by the sequences 5′-G-T-A-A-G-3′ and 5′-A-G-G-A-A-T-3′ at the 5′ and 3′ ends of the introns, respectively. Type 3 corresponds to the boundaries of intron B and contains the direct repeat 5′-A-G-G-T-3′ at both the 5′ and the 3′ end of the intron (all of these sequences are marked by a line above them in Fig. 4).

(*iv*) Two- to five-base pair direct repeats are also found at the 5′ and 3′ limits of a given exon (these nucleotides are marked by dots in Fig. 4).

(*v*) Tracts rich in pyrimidine in the noncoding strand are always found in the intron region preceding the beginning of a new exon (see particularly intron F–exon 6 junction).

(*vi*) The tetranucleotide 5′-T-A-A-G-3′ which is found at the 5′ end of introns D and E is repeated in the reverse order at the 5′ end of exons 4 and 5.

DISCUSSION

Our results indicate that the ov-mRNA has a 45-nucleotide leader sequence (as defined in the Introduction). Leader sequences have been described for several eukaryotic viral mRNAs (for reviews, see refs. 7 and 8). The sequences coding for the leader are most likely not contained in *Eco*RI fragment "b" and should therefore be located upstream from *Eco*RI site 4 (Fig. 1*b*). Yet to be established is the precise location of the sequences coding for the leader and whether they are further split as is the case of adenovirus late mRNAs (15–19). Depending upon the genome position of the sequences for the leader relative to exon 1, our previous minimum estimate of 6000 nucleotides (4) (the distance between the beginning of exon 1 and the end of exon 7, see Fig. 1*b*) for a possible ov-mRNA precursor might now need to be considerably revised. In fact, we have evidence for RNA molecules as large as 10,000 nucleotides containing transcripts of both intron and exon sequences. Although the conalbumin, ovomucoid, and lysozyme genes, which are also split, are not encoded within the 22.4-kilobase genome segment that we have analyzed (see Fig. 1) (M. Cochet, A. Krust, P. Gerlinger, J. L. Mandel, M. LeMeur, and P. Chambon, unpublished results), it cannot be excluded that the ov-mRNA leader sequence is shared by messengers for these or other proteins. However, such a possibility is unlikely because ov-mRNA appears to be the only oviduct mRNA that hybridizes to a DNA probe specific for the leader sequence (unpublished results).

From our previous studies we have concluded that, in contrast to the immunoglobulin case (20–22), there is no rearrangement for the ovalbumin region between *Eco*RI sites 1 and 4 (Fig. 1*b*) during oviduct differentiation. Translocation of the sequences coding for the leader remains a possibility. However, the presence of the *Xba* I site (Figs. 2*b* and 3) in *Eco*RI fragment "b" in both erythrocyte and oviduct DNAs (J. P. LePennec and P. Chambon, unpublished data) reinforces our conclusion that the sequences coding for the leader are not in contiguity with exon 1 in either cell type.

The discovery of the split gene organization has led to the proposal that primary transcripts of these genes could be longer than mRNA and comprise transcripts of both exons and introns. This has been shown to be true for the mouse β-globin gene (23). Although no precursor to ov-mRNA has been demonstrated, we have evidence for just such large ovalbumin gene transcripts (see above). It has been suggested (18) that processing of such precursors could involve the formation of base-paired structures

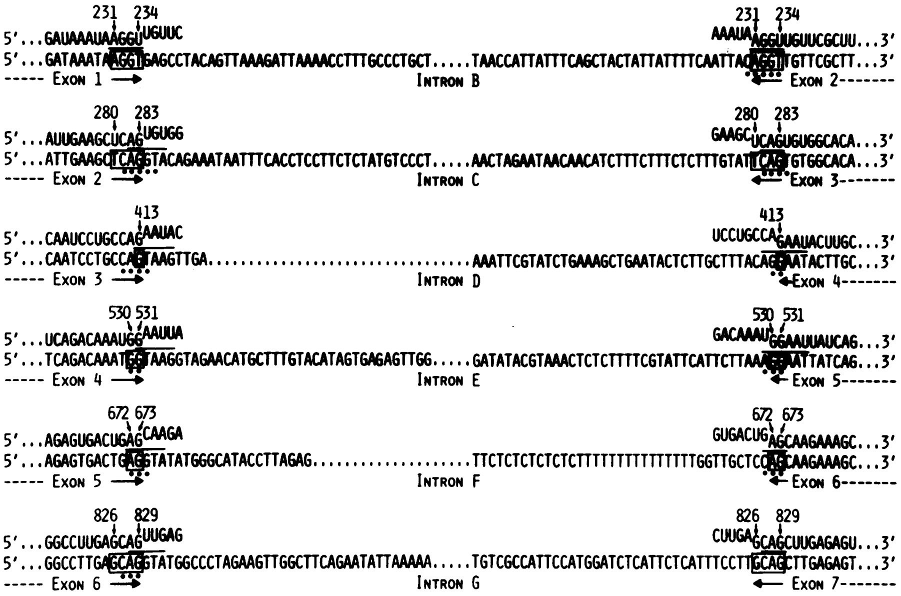
231 234
5'...GAUAAAUAAGGU UGUUC
5'...GATAAATAAGGTGAGCCTACAGTTAAAGATTAAAACCTTTGCCCTGCT.....TAACCATTATTTCAGCTACTATTATTTTCAATTACAGGTTGTTCGCTT...3'
----- Exon 1 →
Intron B
231 234
AAAUA AGGUUGUUCGCUU...3'
← Exon 2-------
280 283
5'...AUUGAAGCUCAG UGUGG
5'...ATTGAAGCTCAGGTACAGAAATAATTTCACCTCCTTCTCTATGTCCCT.....AACTAGAATAACAACATCTTTCTTTCTCTTTGTATTCAGTGTGGCACA...3'
----- Exon 2 →
Intron C
280 283
GAAGC UCAGUGUGGCACA...3'
← Exon 3-------
413
5'...CAAUCCUGCCAG AAUAC
5'...CAATCCTGCCAGTAAGTTGA..............................AAATTCGTATCTGAAAGCTGAATACTCTTGCTTTACAGGAATACTTGC...3'
----- Exon 3 →
Intron D
413
UCCUGCCA GAAUACUUGC...3'
← Exon 4-------
530 531
5'...UCAGACAAAUGG AAUUA
5'...TCAGACAAATGGTAAGGTAGAACATGCTTTGTACATAGTGAGAGTTGG.....GATATACGTAAACTCTCTTTTCGTATTCATTCTTAAAGGAATTATCAG...3'
----- Exon 4 →
Intron E
530 531
GACAAAU GGAAUUAUCAG...3'
← Exon 5-------
672 673
5'...AGAGUGACUGAG CAAGA
5'...AGAGTGACTGAGGTATATGGGCATACCTTAGAG....................TTCTCTCTCTCTCTTTTTTTTTTTTTTTGGTTGCTCCAGCAAGAAAGC...3'
----- Exon 5 →
Intron F
672 673
GUGACUG AGCAAGAAAGC...3'
← Exon 6-------
826 829
5'...GGCCUUGAGCAG UUGAG
5'...GGCCTTGAGCAGGTATGGCCCTAGAAGTTGGCTTCAGAATATTAAAAA.....TGTCGCCATTCCATGGATCTCATTCTCATTTCCTTGCAGCTTGAGAGT...3'
----- Exon 6 →
Intron G
826 829
CUUGA GCAGCUUGAGAGU...3'
← Exon 7-------

FIG. 4. DNA sequences at exon–intron boundaries. The mRNA sequence was derived from the sequence of the ds-cDNA (Fig. 2*a*). Numbering of the nucleotides from the 5′ end of mRNA was as in Fig. 3. The DNA sequence of the noncoding strand at exon–intron boundaries was established as indicated in *Materials and Methods* and Fig. 2 *b–d*. The maximum possible extent of the exons is indicated by the horizontal arrows. Beyond these points, the DNA and RNA sequences diverge and are separated. The numbers define the messenger nucleotides that could be encoded for on either side of the introns. Sequences of nucleotides that are repeated directly at both ends of a given intron are boxed. Sequences of nucleotides that are repeated directly at both ends of a given exon are dotted. Those sequences at the extremities of the different introns that are used to define three types of ovalbumin introns are shown under an unbroken line.

between the opposite ends of an intron transcript to bring into close proximity the ends of transcripts of two neighboring exons with looping out of the intron transcripts. Our sequence data are not compatible with such a model. However, we cannot exclude that more complicated types of folding of the primary transcript could play an important role in an excision–ligation mechanism.

The most striking feature observed from comparison of the exon–intron boundaries of the ovalbumin gene is that not one can be uniquely defined because of the direct repeats described in *Results* (boxed nucleotides in Fig. 4). This allows splicing to occur *a priori* in several different ways while still generating the same spliced product. Taking into account this uncertainty, we have indicated in Fig. 1 *b* and *c* which nucleotides of the ov-mRNA could be encoded by the seven ovalbumin exons. Because the very 3′ end of exon 7 has not been sequenced, we cannot exclude at present that the extreme 3′-terminal nucleotides of ov-mRNA are not encoded for by exon 7. However, such a possibility is unlikely from our previous results (2).

The introns of the ovalbumin gene may be divided into three types depending on the sequences present at their extremities. These types are themselves very closely related and the sequences at the extremities of introns B–G have been aligned in Fig. 5 to emphasize their similarities. It appears that all of the 5′ extremities of the ovalbumin introns can be derived from the sequence 5′-T-C-A-G-G-T-A-3′ with a few base changes and similarly the 3′ intron extremities from the sequence 5′-T-X-C-A-G-G-3′ (Fig. 5). When this alignment is done, it becomes apparent that, in all cases, common excision–ligation points could be defined (broken lines in Fig. 5; see Fig. 4 for mRNA sequences). It is striking that in all cases the dinucleotide at the

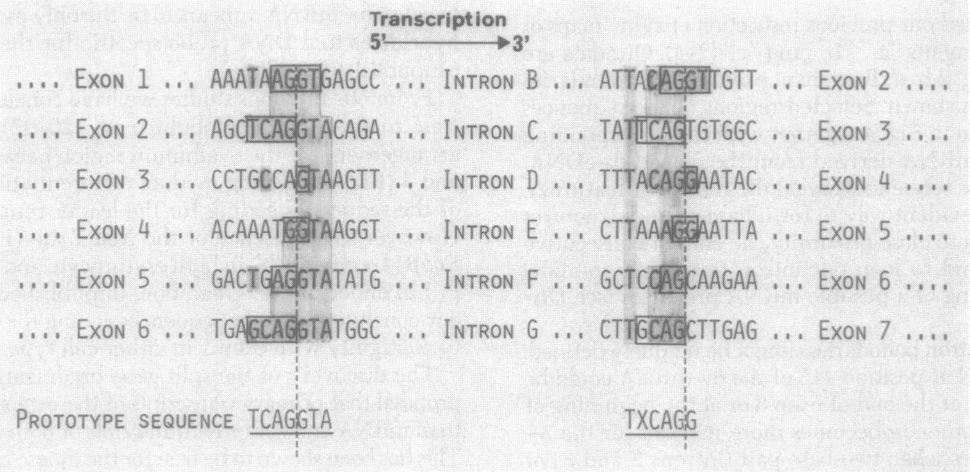
Transcription
5′ → 3′
.... Exon 1 ... AAATAAGGTGAGCC ... Intron B ... ATTACAGGTTGTT ... Exon 2 ....
.... Exon 2 ... AGCTCAGGTACAGA ... Intron C ... TATTCAGTGTGGC ... Exon 3 ....
.... Exon 3 ... CCTGCCAGTAAGTT ... Intron D ... TTTACAGGAATAC ... Exon 4 ....
.... Exon 4 ... ACAAATGGTAAGGT ... Intron E ... CTTAAAGGAATTA ... Exon 5 ....
.... Exon 5 ... GACTGAGGTATATG ... Intron F ... GCTCCAGCAAGAA ... Exon 6 ....
.... Exon 6 ... TGAGCAGGTATGGC ... Intron G ... CTTGCAGCTTGAG ... Exon 7 ....
Prototype sequence TCAGGTA
TXCAGG

FIG. 5. Comparison of the DNA sequences at the ovalbumin gene exon–intron boundaries. Sequences from Fig. 4 have been aligned in order to stress their common features. Boxed nucleotides represent direct repeats as in Fig. 4. The vertical broken line shows how the excision–ligation events could occur in all cases at unique positions with respect to the invariant dinucleotides G-T and A-G (see text). The shadowing stresses the similarities between the individual sequences and the proposed prototype sequences (see text).

5′ end of the introns thus defined is G-T, whereas it is always A-G at their 3′ end.

Comparison of our sequence data with all of the exon–intron junctions of viral and other highly eukaryotic genes sequenced to date, reveals that all of them may be derived from the prototype sequences shown in Fig. 5. This holds true for: (*i*) small and large introns of rabbit and mouse β-globin (J. Van den Berg, A. Van Ooyen, N. Mantei, A. Schambock, R. Flavell, and C. Weissmann, personal communication); (*ii*) small introns of two λI and one λII immunoglobulin light chains and a long intron in one λI immunoglobulin light chain (ref. 5; O. Bernard, N. Hozumi and S. Tonegawa, personal communication); and (*iii*) introns of several late and early simian virus 40 genes (ref. 24; P. K. Ghosh, V. B. Reddy, J. Swinscoe, P. Lebowitz, and S. Weissman, personal communication). In third case there are striking similarities with the different types of ovalbumin intron extremities: (*i*) the intron extremities of the early T antigen gene and of the early t antigen gene are similar to those of ovalbumin type 1 intron extremities; (*ii*) the intron extremities of one late 19S RNA gene and of the late 16S gene are similar to those of ovalbumin type 2 intron extremities; and (*iii*) the intron extremities of two other late 19S RNA genes are identical to those characteristic of ovalbumin type 3 intron extremities.

In all of the above cases, as for the ovalbumin, the splicing point is not uniquely defined by the nucleotide sequence at the boundaries. However, all of the sequences of the exon–intron junctions of these genes can be aligned as has been done for the ovalbumin gene in Fig. 5. Again, this alignment allows definition, for all of these genes, of unique common excision–ligation points as shown in Fig. 5 for the ovalbumin, and again particularly noteworthy is the invariance of dinucletodies G-T (in all cases) and A-G (with one exception) at the 5′ and 3′ extremities of the introns, respectively. It is thus possible that splicing may occur at unique points even though at first sight the nucleotide sequences of the transcript at the boundaries do not allow such a conclusion. Whether the splicing point is in fact unique, whether the above dinucleotides could be part of the site(s) recognized by the enzyme machinery responsible for the necessary accuracy of the excision–ligation events, and whether secondary and tertiary foldings of the intron transcripts may bring them in close proximity are at present unknown. In this respect it it interesting to note that, in all cases, as for ovalbumin, the 3′ end of the introns is preceded by a pyrimidine-rich tract. It should be noted that the above dinucleotide are not found at the extremities of the yeast tRNA introns (25, 26), in which cases the intron extremities do not appear to be derived from the prototype sequences shown in Fig. 5, although in these cases the excision-ligation points are also not uniquely defined by the sequences.

Direct repeats of nine and five base pairs have been found at the extremities of insertion sequences IS1 (27, 28) and other translocatable elements (29), respectively, when integrated into a host genome. The nine-base pair repeat of IS1 differs from one insertion to another, but the different repeats appear to be somewhat related and may be derived from a unique sequence (28). It is tempting to speculate that the extremities of introns may have evolved from an analogous common direct repeat. That this might have been the case is hinted at by the existence of a four-base-pair direct repeat 5′-C-A-G-G-3′ in our prototype sequences (Fig. 5) and by the direct repeats found at the extremities of a given exon (see Fig. 4, dotted lines). These analogies suggest the possibility that the mechanisms responsible for the appearance of introns might be related to those involved in the integration of insertion elements.

We thank Mr. E. Taubert and Mr. J. M. Garnier for excellent technical assistance. We are grateful to Drs. V. Pirotta, P. Tiollais, and J. Sussenbach for gifts of restriction enzymes and to Dr. F. Rougeon for a gift of polynucleotide kinase T4. We thank Drs. G. Brownlee (Cambridge) and J. P. LePennec (Strasbourg) for communication before publication of ov-mRNA sequence and restriction enzyme maps, respectively. We are grateful to Dr. P. Kourilsky for useful discussions. This work was supported by grants from the Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique (ATP "Structure primaire des Acides Nucléiques," décision 3558) and the Fondation pour la Recherche Médicale Française. R.B. was supported by a European Molecular Biology Organization long-term fellowship.

1. Breathnach, R., Mandel, J. L. & Chambon, P. (1977) *Nature (London)* **270**, 314–319.
2. Garapin, A. C., LePennec, J. P., Roskam, W., Perrin, F., Cami, B., Krust, A., Breathnach, R., Chambon, P. & Kourilsky, P. (1978) *Nature (London)* **273**, 349–354.
3. Garapin, A. C., Cami, B., Roskam, W., Kourilsky, P., LePennec, J. P., Perrin, F., Gerlinger, P., Cochet, M. & Chambon, P. (1978) *Cell* **14**, 629–639.
4. Mandel, J. L., Breathnach, R., Gerlinger, P., LeMeur, M., Gannon, F. & Chambon, P. (1978) *Cell* **14**, 641–653.
5. Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O. & Gilbert, W. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1485–1489.
6. McReynolds, L., O'Malley, B. W., Nisbet, A. D., Fothergill, J. E., Givol, D., Fields, S., Robertson, M. & Brownlee, G. G. (1978) *Nature (London)* **273**, 723–728.
7. Sambrook, J. (1977) *Nature (London)* **268**, 101–104.
8. Chambon, P. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 1211–1236.
9. Humphries, P., Cochet, M., Krust, A., Gerlinger, P., Kourilsky, P. & Chambon, P. (1977) *Nucleic Acids Res.* **4**, 2389–2406.
10. Smith, H. O. & Birnstiel, M. L. (1976) *Nucleic Acids Res.* **3**, 2387–2398.
11. Efstratiadis, A., Vournakis, J., Donis-Keller, H., Chaconas, G. & Kafatos, F. (1977) *Nucleic Acids Res.* **4**, 4165–4172.
12. Maniatis, T., Kee, S. G., Efstratiadis, A. & Kafatos, F. C. (1976) *Cell* **8**, 163–182.
13. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
14. Palmiter, R. D., Gagnon, J. & Walsch, K. A. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 94–98.
15. Berget, S. M., Moore, C. & Sharpe, P. A. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3171–3175.
16. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. (1977) *Cell* **12**, 1–8.
17. Dunn, A. R. & Hassel, J. A. (1977) *Nature (London)* **259**, 596–598.
18. Klessig, D. F. (1977) *Cell* **12**, 9–21.
19. Lewis, J. B., Anderson, C. W. & Atkins, J. F. (1977) *Cell* **12**, 37–44.
20. Hozumi, N. & Tonegawa, S. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 3628–3632.
21. Rabbitts, T. H. & Forster, A. (1978) *Cell* **13**, 319–327.
22. Brack, C., Hirama, M., Lenhard-Schuller, R. & Tonegawa, S. (1978) *Cell*, in press.
23. Tilgham, S., Curtis, P., Tiemeier, D., Leder, P. & Weissmann, C. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1309–1313.
24. Reddy, V. B. Thimmappaya, B., Dhar, R., Subramanian, K. N., Zain, B. S., Pan, J., Ghosh, P. K., Celma, M. L. & Weissman, S. M. (1978) *Science* **200**, 494–502.
25. Goodman, H. M., Olson, M. V. & Hall, B. D. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5453–5457.
26. Valenzuela, P., Venegas, A., Weinberg, F., Bishop, R. & Rutter, W. J. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 190–194.
27. Calos, M. P., Johnsrud, L. & Miller, J. H. (1978) *Cell* **13**, 411–418.
28. Grindley, N. D. F. (1978) *Cell* **13**, 419–426.
29. Rosenberg, M., Court, D., Wulff, D. L., Shimatake, H. & Brady, C. (1978) *Nature (London)*, **274**, 213–214.