Primary Sequence of Ovomucoid Messenger RNA as Determined from Cloned Complementary DNA

JAMES F. CATTERALL, JOSEPH P. STEIN, PAULA KRISTO, ANTHONY R. MEANS, and BERT W. O’MALLEY

Department of Cell Biology, Baylor College of Medicine, Houston, Texas 77030

ABSTRACT Ovomucoid messenger RNA (mRNAom) comprises ~8% of the total mRNA in the estrogen-stimulated oviduct. The recombinant plasmid pOM100 contained DNA complementary to the 3’ end of mRNAom. DNA complementary to the 5’ end of mRNAom was obtained from a partially purified preparation of mRNAom by polymerization by reverse transcriptase in the presence of a restriction fragment primer from pOM100. The complementary DNA mixture was amplified by molecular cloning using poly dG/dC tailing to form recombinant bacterial plasmids. Recombinant plasmids containing ovomucoid DNA sequences were selected by in situ hybridization to 32P-labeled pOM100 fragments. The longest plasmid containing ovomucoid DNA sequences was designated pOM502. The complete DNA sequence of both pOM100 and pOM502 was determined. The two plasmids appear to contain sequences complementary to the entire length of mRNAom. The nucleic acid sequence agrees with the known amino acid sequences for both ovomucoid and its N-terminal signal peptide. Highly homologous sequences occur in two regions that coincide with structural domains of the protein. Comparison of the sequence of mRNAom with that for other eucaryotic mRNAs allowed identification of possible functional regions in the mRNA molecule.

The synthesis of ovomucoid and ovalbumin is regulated by steroid hormones in the chick oviduct (16, 26). The genes that code for these two proteins offer an attractive model system for the study of coordinate expression of unlinked genes in a steroid-hormone target tissue.

We have previously reported that the ovomucoid gene has a complex structure containing seven intervening sequences (9, 21). Messenger RNA sequences are required for the study of the fine structure of complex eucaryotic genes. Previous knowledge of mRNA sequences allowed the identification of the cleavage points at which splicing of primary transcripts must occur (6, 8, 19). Features of eucaryotic mRNA, such as 5’ leader sequences, initiation signals, 3’ noncoding regions, and possible secondary structures have been studied by nucleic acid sequencing (2, 11, 23). Comparison of mRNA sequences from various sources have led to the identification of conserved regions which may have structural or functional significance (2, 20, 27, 29). Sequences of coding regions have elucidated unknown protein sequences (23) and will allow correlation of conservation of nucleic acid sequences with protein structural and functional domains.

To determine the nucleotide sequence of the ovomucoid structural gene, we have synthesized and cloned DNA complementary to essentially the entire ovomucoid mRNA (mRNAom). The sequence was contained within the plasmids pOM100, (33) which included the 3’ end and pOM502, which contained the 5’ sequence. The entire nucleic acid sequence corresponding to mRNAom is contained in the present manuscript and provides an interesting comparison with the amino acid sequence of ovomucoid (17). The sequence is also compared with sequences of other eucaryotic mRNAs, particularly that for ovalbumin, in an effort to identify common primary and secondary structural features.

MATERIALS AND METHODS

Materials

Restriction enzymes Hae III, Hinc II, Hpa II, EcoRI, Alu I, Bam HI, and Sau IIIA were obtained from Bethesda Research Laboratories, Rockville, Md. Restriction enzymes Hinf I, Hha I, Hph I, Mbo II, Psi I, Kpn I, and Ava II were purchased from New England Biolabs, Inc., Beverly, Mass. S, nuclease was obtained from Miles Laboratories, Inc., Elkhart, Ind. Bacterial alkaline phosphatase was purchased from Worthington Biochemical Corp., Freehold, N. J., and T4 polynucleotide kinase from Boehringer Mannheim Biochemicals, Indianapolis, Ind. AcA34 was purchased from LKB Instruments, Inc., Rockville, Md. [γ-32P]ATP (3,000 Ci/mmol) was purchased from Amersham Corp., Arlington Heights, Ill. Hydrazine was purchased from Pierce Chemical Co., Rockford, Ill., and carbonate-free NaOH from J. T. Baker Chemical Co., Phillipsburg, N. J.
Internally primed cDNA synthesis

The recombinant plasmid pOM100 (600 µg; 33) was digested with 400 U Pst I (1 U will digest 1 µg of DNA/µl at 37°C) for 15 h at 30°C in 1.5 ml of reaction volume. The 204 bp Pst I fragment (33) was isolated by electrophoresis in a 4.5% polyacrylamide slab gel. The gel was prepared in a total volume of 150 ml with a 1:20 ratio of N,N'-methylenebisacrylamide to acrylamide, 200 µl of N,N',N'-tetramethylethylenediamine, and 0.13 g of ammonium persulfate in 40 mM Tris-acetate, pH 8.0, 16 mM NaCl, 1 mM EDTA. The fragment was eluted from a gel slice according to the method of Maxam and Gilbert (22).

To remove residual polyacrylamide, an 11.5 x 0.7-cm AcA34 gel filtration column was prepared in 10 mM Tris-HCl, pH 7.6, 100 mM NaCl. The column was calibrated and equilibrated with transfer RNA and undigested recombinant plasmid DNA. Fractions of 0.6 ml were collected and assayed by analytical polyacrylamide gel electrophoresis. Fractions containing the 204 bp Pst I fragment were pooled.

The purified fragment (0.5 µg) was denatured at 100°C for 1 min and quickcooled in ice. A 30-fold molar excess of partially purified mRNA was added (33) in a final reaction volume of 250 µl containing 10 mM Tris-HCl, pH 7.4, 100 mM NaCl, and 2 mM EDTA. The hybridization reaction was terminated after 1 min at 68°C (Rt = 2.5 x 10³) by freezing at -70°C. The primer-template complex was precipitated with the nonspecific RNA in the mixture with 2.5 vol of 95% ethanol.

The synthesis of DNA complementary to the 5' end of mRNA was by reverse transcriptase (242 U/ml) was performed as previously described (36). Under the conditions described above, complementary cDNA was synthesized at oligo dT independent. Synthesis of double-stranded cDNA, and cleavage of the terminal hairpin loop were carried out as described previously (33).

Preparation and Amplification of Recombinant Plasmids

Pst I cleaved pBR322 (4) was tailed with G residues and ds cDNA, with terminal poly (dC) residues in the presence of terminal transferase under conditions described previously (33). The two species were mixed in a 1:1 M ratio and annealed after heating at 70°C in 10 mM Tris, pH 7.4, 100 mM NaCl, 1 mM EDTA.

The recombinant plasmid mixture (1.5 µg/ml) was mixed with 2 vol of CaCl₂-treated recipient cells (Escherichia coli K strain RR1; 4). After incubation on ice for 60 min, 15-µl aliquots were spread into L-agar plates containing 25 µg/ml tetracycline (Tc) and incubated at 37°C. Tc-resistant colonies were picked and transferred to a fresh Tc plate.

Selection of cDNA om-containing Recombinants

Tc-resistant colonies were transferred to a Millipore filter disk (Millipore Corp., Bedford, Mass.) and lyed in situ by the method of Grunstein and Hogness (13). The filter was hybridized in the presence of 20 µl labeled pOM 204 (bp) fragment after treatment in Denhardt's solution (5, 10). Positive colonies were grown separately, and plasmid DNA was prepared by a modification of the method of Katz et al. (18). Plasmid DNA was incubated with restriction enzymes Pst I or HhaI and the digestion products were separated on a 6% polyacrylamide slab gel. The purified fragment (0.5 µg) was denatured at 100°C for 1 min and quickcooled. The complementary strands of the DNA molecule were separated by electrophoresis through a 20% polyacrylamide/7 M urea thin gels.

Preparation of End-labeled Fragments

Routinely, 40 µg of plasmid DNA was digested with an appropriate restriction endonuclease at 1 µg/µl DNA for 2-3 h at 37°C in a 200 µl reaction mixture. After precipitation with 2.5 vol of ethanol (95%) in the presence of 0.3 M NaOAc, the DNA was dissolved in 40 µl of H₂O for 5' end labeling, or 31 µl of 10% sucrose, 0.05% bromophenol blue, 0.25% xylene cyanole FF, boiled for 3 min, chilled, and loaded onto 8 and 20% polyacrylamide/7 M urea thin gels. Length of each electrophoresis run was determined by fragment length and the region of interest to be sequenced.

RESULTS

Recombinant Plasmids Containing cDNA om

Preparation of the recombinant plasmid pOM100, which contains cDNA synthesized from mRNA, was described previously (33). This plasmid contained a 650 bp DNA insert, but lacked ~150 by corresponding to the 5' end of mRNA. To obtain DNA sequences containing the 5' end of the mRNA sequence, we have synthesized and cloned another DNA molecule with a fragment of pOM100 as a specific primer.

Plasmid pOM100 was digested with EndoR Pst I and the 204 bp central fragment was purified from an acrylamide gel slice. The purified fragment was then hybridized with a partially purified preparation of mRNA-DNA complementary to the 5' end of mRNA. To obtain DNA sequences containing the 5' end of the mRNA sequence, we have synthesized and cloned another DNA molecule with a fragment of pOM100 as a specific primer.
was used to transform *E. coli* K strain RRI (4), as described in Materials and Methods. All bacterial transfers were carried out in a certified laminar flow hood in a P-3 physical containment facility. (These experiments were approved at P-3, EKI containment in accordance with the Revised Guidelines for Recombinant DNA Research, published by the National Institutes of Health.)

Tc-resistant transformants, of which 49 were obtained, were transferred directly to a nitrocellulose filter and assayed by *in situ* hybridization. In this study, 30 of 49 Tc-resistant clones contained inserted DNA complementary to a [³²P]pOM100-*Pst* hybridization probe. The 19 clones that did not hybridize with the pOM100 probe probably represented uncut pBR322. It has been shown that molecules tagged with dGTP residues can reanneal in an intramolecular reaction to form stable monomers (A. Dugaiczky, personal communication). These 19 clones were not characterized further to distinguish these possibilities. The recombinant plasmids were screened for insert size by agarose gel electrophoresis. The largest plasmid, designated pOM502, was further characterized by restriction mapping and Southern blot (30) analysis, as previously described for pOM100 (33). Fig. 1 shows the positions and extent of the DNA inserts of plasmids pOM100 and pOM502. Plasmid pOM502 contains a DNA*om* insert of 538 nucleotides, excluding GC tails of 11 bp at the 5' end (defined as the 5' end of the mRNA sequence) and 16 bp at the 3' end. Together, the two plasmids contain 821 nucleotides of cDNA*om*.

**Determination of Nucleotide Sequence**

The entire nucleotide sequences of the two cDNA*om* clones described above were determined by the chemical modification method of Maxam and Gilbert (22). Restriction fragments used in the sequence determinations are shown diagrammatically in Fig. 2. 11 restriction sites were used as labeling sites. The various fragments overlap in several regions, allowing two independent sequence determinations in many areas. Where practical, the sequence was also determined from the complementary strand by 3' end labeling or by judicious selection of an alternative 5' labeling site. A potential hazard in the design of sequencing strategy is the presence of small fragments between closely spaced restriction sites that have been used for labeling or recutting. Thus, all sites of labeling and recutting were independently sequenced within other restriction fragments to eliminate this possibility.

The sequence of the translated region (54-686) was further checked by comparison with the amino acid sequence (17, 34). The 5' noncoding region was sequenced mainly on one strand from unambiguous gel patterns. Sequence data from genomic clones, which have been shown to contain no intervening sequences in the 5' noncoding region, are in complete agreement with the data from the cDNA clone.

The sequence of the 3' noncoding region was reported by Buell et al. (7). Our initial data, derived from experiments on only one strand in the 3' noncoding region, revealed 18 discrepancies with their published sequence. To resolve this disagreement, the sequences of both complementary strands were determined. The data (Fig. 3) clearly show differences in the sequence of pOM100 and that published by Buell et al. (17). The number of differences (18) makes it unlikely that all arise from variations between the two clones.

A computer search (31, 32) of the sequence showed the positions of all known restriction enzyme recognition sites (Table I). Table I includes restriction sites throughout the mRNA sequence, not only those used in the sequence analysis. Some sites listed in Table I were determined on the basis of sequence alone, although most have been shown to occur by actual cleavage studies as well.

**5' Noncoding Sequence**

The 5' noncoding region is 53 nucleotides in length. The 5' terminal nucleotide in the mRNA*om* sequence as determined from the cDNA is adenine. This is consistent with the capping sites of other eucaryotic mRNAs where A often follows the 7-methyl G cap (20). Because pOM502 was made as a short reverse transcription product from a cloned primer fragment, the cDNA product may include sequences complementary to the 5' terminus of mRNA*om*. The length of the complete cDNA sequence (821 nucleotides) is in good agreement with the

---

**FIGURE 2** Diagram of restriction fragments used in sequencing. Restriction sites for which positions are shown are only those from which sequence information was derived. The number of sites does not necessarily reflect the total number in the sequence (see Table I). Arrows indicate the extent of sequence determined from labeling sites listed at left. The direction of the arrow indicates which strand of the cDNA was sequenced; arrows pointing away from the labeling site indicate that the 5' strand was labeled and sequenced, and arrows pointing toward the labeling site indicate that the 3' strand was labeled and sequenced. Broken lines indicate portions of the sequence that were not read in a particular experiment (usually 80-cm long gels were used to obtain specific regions of the sequence remote from the labeling site.

---

**FIGURE 1** cDNA clones derived from mRNA*om* cDNA*om* was synthesized from mRNA*om* (top) with a primer fragment from pOM100 (middle) and inserted into plasmid vector pBR322 (see Materials and Methods). Plasmid pOM502 (bottom) contained DNA complementary to the 5' end of mRNA.
FIGURE 3  Sequence derived from the 3' terminal portion of mRNA<sub>om</sub>. The sequences shown in a, b, and c were derived from a 169 bp fragment of an RI digest of pOM100 recut with Hinc II. The fragment was 5' end labeled. The sequences shown in d and e were derived from a 195 bp fragment of a Hind I digest of pOM100 recut with Hha I. This fragment was 3' end labeled. Therefore, all sequences shown correspond to the anti-mRNA strand. (a) This sequence, derived from a 10% polyacrylamide/7 M urea gel, corresponds to positions 604-664 in Fig. 4. (b) This sequence, derived from a 10% polyacrylamide/7 M urea gel, corresponds to positions 661-690 in Fig. 4. (c) This sequence, derived from a 20% polyacrylamide/7 M urea gel, corresponds to positions 682-703 in Fig. 4. (d) This sequence, derived from a 20% polyacrylamide/7 M urea gel, corresponds to positions 732 to 778 in Fig. 4. (e) This sequence, derived from a 10% polyacrylamide/7 M urea gel, corresponds to positions 775-821 in Fig. 4, then into the poly A tail (poly T in this, the antistrand).

Restriction Endonuclease Recognition Sites in pOM100 and pOM502

| Enzyme*  | Position$ | Sequence |
|----------|-----------|----------|
| Hae III  | 56, 392, 473 | GG/CC |
| Hinc II  | 480, 531 | GTY/RAC |
| Hind I   | 226, 421, 653, 728 | G/ANTC |
| Hpa II   | 18, 37 | C/CGG |
| Hpa I    | 424, 706, 746 | GGTGA or TCACC |
| EcoRI    | 702 | G/AATTC |
| Alu I    | 694, 774, 779 | AG/CT |
| Barn HI  | 714 | G/GATCC |
| Bal I    | 55§ | TGGCCA |
| Sau IIIa | 715 | /GATC |
| Mbo II   | 70, 570 | GAAGA or TCTTC |
| Pst I    | 29, 137, 332, 536 | CTGCA/G |
| Sst I    | 773§ | GAGCT/C |
| Kpn I    | 216 | GGTAC/C |
| Ava II   | 197, 377 | G/G (A/T)CC |
| Asu I    | 197, 377, 391, 472§ | G/GNCC |

Restriction endonuclease recognition sites in pOM100 and pOM502. The positions of recognition sites were determined by computer analysis (31, 32) of the complete mRNA<sub>om</sub> sequence. All positions, except where indicated, have been detected by actual enzyme cleavage studies as well. The cleavage site, when within the recognition sequence, is designated by / . The cleavage point of Bal I is unknown. Y, pyrimidine; R, purine.

* Numbers correspond to the first nucleotide in the recognition sequence and, therefore, do not necessarily correspond to the cleavage site.
$ Enzymes tested for which no site was present include: Hind III, Hae II, Hha I, Hpa I, EcoRI, Ava I, Bcl I, Bgl II, Taq I, Taq I, Pvu II, Sst I and III, Sal I, Xho I, and Xba I.
§ These positions were determined from the computer search of the sequence only.

The estimated length of mRNA<sub>om</sub> (33). Furthermore, another plasmid, obtained in these experiments, of similar size to pOM502 was shown to terminate at exactly the same position as pOM502 (data not shown). This suggests that reverse transcription of mRNA<sub>om</sub> from the internal primer went to completion. However, we cannot formally exclude the possibility that a few nucleotides were removed during S<sub>n</sub> nuclease cleavage of the covalently closed double-stranded cDNA.

The 5' noncoding sequence is very GC-rich (67.2%) as compared with the total mRNA sequences (52.0%) or chicken DNA (43.5%). This segment is particularly low in U. Only 5 U's occur in the 53 nucleotides preceding the AUG initiator codon.

The initiator AUG is the first AUG triplet in the sequence. It has been suggested that AUG is the only signal sequence in the 5' noncoding region of eucaryotic mRNA (2). Two short regions of complementarity with the 3' end of 18S rRNA occur in this region. One, CACC, occurs immediately before the initiator AUG and is complementary to nucleotides 21-24 from the 18S rRNA 3' end (1). Another sequence, GCAG, which is complementary to nucleotides 14-17 in 18S rRNA (1), is repeated four times at quite regular intervals from the complementary sequence at the AUG. These structural features may be important in properly positioning mRNA<sub>om</sub> on the ribosome during translation.

Translated Region

Ovomucoid has recently been sequenced (17) and the amino acid sequence corresponds to the nucleotide sequence from
positions 135 to 686. The two sequences agree completely, lending credence to the accuracy of each.

Ovomucoid is a secreted protein and, like other secretory proteins, has a hydrophobic signal peptide (3) at its amino terminus. The sequence of this peptide (34) agrees with the nucleic acid sequence from 54–125 in Fig. 4. The entire translated region is 629 nucleotides in length and is terminated by the nonsense codon, UGA.

Codon usage in eucaryotic mRNAs is nonrandom (11, 14, 23). Table II shows codon usage in mRNAom. There are 11 codons of 61 possibilities (excluding nonsense codons) that are not represented in mRNAom. Several amino acids are coded by an especially nonrandom set of synonymous codons, particularly those for Asn and Arg. The only amino acid not represented at least once is tryptophan. 40% of the codons have C in the third position. The distribution of C-terminated codons is identical in both the signal peptide and the mature protein sequence.

Codons containing the dinucleotide CpG are particularly rare, occurring only three times in 211 coding triplets. Without regard to coding frame, 26 CpG dinucleotides occur in mRNAom. The dinucleotide UpA occurs only 23 times in mRNAom, making it the least represented of the 16 possible dinucleotide combinations. However, it occurs 10 times within codons.

The amino acid sequence of ovomucoid revealed three structural domains in the protein (17). Comparative sequencing results from several species indicated that domains I and II had diverged more recently than II and III. Homology between the amino acid sequences of domains I and II supported this theory. Table III shows the homology between the nucleotide sequences of domains I and II supported this theory. Table III shows the homology between the nucleotide sequences of domains I and II.

The nucleotide sequences within domains I and II (positions 126–320 and 321–515 in Fig. 4) are homologous. Even greater homology (85%) exists between the regions 216–263 and 411–458 within domains I and II. Domains II and III, and I and III
chains, which may be required to maintain protein structure or domains. Represent diverging copies of a primordial sequence, mRNA sequences (36%). As might be expected if the sequences with the 3' noncoding region (34%) or with ovalbumin (35%) show less homology, being only slightly above background. Random homology was established by comparing domain I sequences, which is conserved in eucaryotic mRNAs (27), is present only in mRNA om. The position of this conserved sequence at the polyA is similar to its position in other eucaryotic mRNA molecules (27).

### DISCUSSION

Nucleotide sequences of eucaryotic messenger RNA have provided important information regarding mRNA structure (2, 11, 14, 23) and possible functional sites, such as ribosome binding sites (20, 29) and the AAUAAA near the 3' end (27). Secondary structures, functional codons, 5' leader sequences, 3' noncoding regions, and unknown protein sequences have all been elucidated by mRNA sequences. The evaluation of protein function may be reflected in highly conserved coding sequences. Recently, mRNA sequences have been crucial in defining the fine structure of complex eucaryotic genes. Genomic DNA contains both structural and intervening sequences that cannot be distinguished without complete mRNA sequences. The splice points at which RNA processing occurs have been determined by comparison of mRNA sequences with native gene sequences (6, 8, 19). The mRNA om sequence reported here was used as a guide in identifying sequences flanking the 5' and 3' ends of the ovomucoid gene (21).

The recombinant plasmid pOM100 (33) contained ~80% of the mRNA om sequence, including the 3' end. To obtain a copy of the 5' end of the mRNA om sequence, we used a restriction fragment from pOM100 as a specific primer. The primer was used to clone the 5' sequence of mRNA om in the presence of contaminating mRNA species. Cloned complementary DNAs synthesized from oligo dT primers have been shown not to contain DNA complementary to the 5' terminus of the mRNA template (11, 14, 15, 23, 24, 33). In our experience, the longest cDNAs (examined after cloning) vary widely in length (24, 33). Our present findings suggest that it may be possible to transcribe cDNA to the 5' terminus of an mRNA template and maintain full length during cloning procedures. The two longest cDNA inserts obtained in these experiments terminate at an identical nucleotide position in mRNA om, which may be the 5' terminus. Synthesizing short cDNA products from cloned internal primers may provide an advantage in obtaining complete cDNA sequences. However, to confirm this point, direct end labeling and sequencing of mRNA om is required.

The mRNA om sequence can be separated into three segments, the 5' and 3' noncoding regions and the translated region. The 5' noncoding region is 53 nucleotides in length. Ovalbumin mRNA has a 5' noncoding region of similar length (23). The AUG codon at the beginning of the translated region is the only AUG triplet in the 5' noncoding sequence. Whereas recognition of the first AUG codon during initiation of translation is a general feature of eucaryotic mRNAs, the distance between the AUG and the 5' cap varies greatly (20). It has been suggested that the AUG initiator is the only signal sequence in the 5' noncoding region of eucaryotic mRNAs (2). However, some interesting homologies were revealed by comparing ovomucoid and ovalbumin mRNAs. There is an 11-nucleotide sequence surrounding the initiator AUG that is homologous in these two mRNAs. The sequence UCACUAUGGNC occurs in both molecules where N is the only nonhomologous nucleotide. This sequence contains the tetranucleotide CACC, which is complementary to the 3' end of 185 rRNA (1). Another homologous region is the tetranucleotide GCAG, which occurs at -33 and -34 from the AUG in

---

**TABLE II**

| Codon Usage in mRNA om |
|------------------------|
| U | C | A | G |
|---|---|---|---|
| Phe | 4 | 1 | 2 | 8 |
| Leu | 6 | 1 | 0 | 2 |
| Ser | 6 | 0 | 0 | 0 |
| His | 2 | 1 | 2 | 0 |
| Pro | 5 | 0 | 1 | 2 |
| Thr | 1 | 3 | 6 | 1 |
| Ala | 6 | 5 | 7 | 3 |
| Val | 6 | 0 | 10 | 7 |
| Glu | 6 | 3 | 7 | 7 |
| Asp | 5 | 7 | 6 | 3 |

**TABLE III**

| Homology between Sequences of Structural Domains of Ovomucoid |
|-------------------|
| Domains | NA | AA | AA groups |
|---------|---|---|----------|
| I/II | 66 | 49 | 71 |
| II/III | 38 | 35 | 57 |
| I/III | 40 | 36 | 70 |

Percent of homology between structural domains of ovomucoid. Sequences were compared according to the alignment of the structural domains of the protein by Kato et al. (17). NA, nucleic acid sequence; AA, amino acid sequence; AA groups, amino acids grouped according to characteristic side chain: hydrophobic—Phe, Leu, Ile, Met, Val, Pro, Ala, Tyr, Trp; polar—Ser, Thr, Glu, Asn, Cys, Gly; basic—His, Lys, Arg; acidic—Asp, Glu.
ovomucoid and ovalbumin mRNAs, respectively. This sequence is also complementary to the 3' end of rRNA (1). It has been suggested that these regions, which appear to be conserved among several eucaryotic mRNAs, may act to properly position the AUG initiator on the 40S ribosomal subunit (29). There is no other apparent homology between ovomucoid and ovomucoid mRNA in their 5' noncoding regions.

Ovalbumin mRNA may have a stable hairpin-loop structure at its 5' end (23, 29). This structure may serve to reduce the cap—AUG distance and facilitate ribosome binding or initiation of translation. No such stable structure can be drawn for the 5' end of mRNAom. However, it is interesting that GCAG, which is present in both ovalbumin and mRNAom and has 18S rRNA binding potential, is repeated four times in the 5' noncoding region of mRNAom at regular intervals from the homologous sequence surrounding the AUG. The cap—AUG distance in mRNAom could be reduced by interaction of the 3' end of 18S rRNA with these regularly repeated sequences.

The internal homology in the translated region reflects the known protein structure of ovomucoid. Three homologous domains were identified from the amino acid sequence. The higher level of homology seen in the nucleotide sequence supports the gene duplication theory of Kato et al. (17). The very strong homology between two regions of domains I and II (216-263 and 411-458) may be indicative of a functional role for the amino acids coded in these regions. This possibility is difficult to assess because the in vivo function of ovomucoid is unknown. The active site for in vitro trypsin inhibition by ovomucoid lies in domain I outside the highly conserved region (17).

The use of synonymous codons in the translated region is nonrandom. Selective use of isoaccepting tRNA may provide a basis for developmental control of some proteins. This mechanism appears not to occur in the ovuduct, as ovomucoid and ovalbumin are coordinately expressed but have different isoaccepting tRNA requirements. The mRNAom, translated sequence is rich in C-terminated codons. 40% of the 211 triplets end in C. Pyrimidine-terminated codons are favored over purine-terminated codons 66-34%. Of the four nonviral eucaryotic mRNAs sequenced thus far, those for α-globin (14) and ovomucoid strongly favor pyrimidine-terminated codons (64 and 66%, respectively), and β-globin (11) and ovalbumin (23) are relatively neutral in this respect (56 and 53%, respectively).

Among the pyrimidine-restricted codons (12), those ending in C predominate (63%) in mRNAom as well. Fitch (12) suggested that preferential use of C in the third position of codons selects for nonwobble base pairing during translation, resulting in a lower error rate. At the active site for trypsin inhibition by ovomucoid, three pyrimidine-restricted codons occur, all of which end in C. In the highly conserved regions of protein domains I and II discussed above, C-terminated codons predominate among the pyrimidine-restricted codons, but they are not used exclusively. In the absence of a well-characterized in vivo function for the protein, it is difficult to assess the importance of this phenomenon to the maintenance of protein function.

Codons containing the dinucleotide CpG are particularly rare. Only three of 211 triplets contain CpG. If the mRNA is taken as a whole without regard to reading frame, 26 CpG dinucleotides occur. This is significantly below the 51 expected by chance in a random 821 nucleotide sequence. Other eucaryotic mRNAs (11, 14, 23) have been shown to contain few CpG dinucleotides. The reason for this is unknown but it has been suggested that methylation of cytosine residues at CpG makes this a hot spot for mutation, causing low CpG levels (14, 25).

The sequences presented here have been independently tested by comparison with amino acid sequences (17, 34). The protein and nucleic acids sequences agree completely (M. Laskowski, personal communication). Noncoding sequences were determined from two independent sources, the cDNA clones pOM502 (5') and pOM100 (3') and the ovomucoid native gene clones (9, 21). Throughout the sequence, all restriction sites used for 5' end labeling or secondary cleavage were determined from overlapping fragments. This precaution allowed us to prove that sequences found on either side of a labeling or recutting site are contiguous and eliminates the possibility of missing a small fragment between two closely spaced restriction sites. Despite the apparent accuracy indicated by the agreement between the nucleic acid and protein sequences and the precautions taken to maintain sequencing accuracy, we cannot completely exclude the possibility of an error.

The 3' noncoding region and terminal 75 nucleotides of coding sequences were sequenced from the unique EcoRI site, and apparently one HindI site, by Buell et al. (7). Our sequence of this region, derived from both 5' and 3' labelings of the EcoRI site and HindI sites (positions 653 and 728), differs from their published sequence at 18 positions (see Fig. 3). Whereas an occasional variation between cloned sequences may be possible, we suspect that the large number of discrepancies in these sequences is caused by technical differences. We attempted to further validate our sequence by carrying out determinations on each of the complementary strands.

The mRNAom sequence has been used to identify possible regulatory sequences that flank the ends of the native gene (21). The sequences at the 14 "junctional sequences" within the ovomucoid gene are currently being determined. It is hoped that comparison of these sequences with those of ovalbumin (6, 8) and other eucaryotic genes (19, 35) will lead to the isolation or direct synthesis of an "average" junctional splicing site. This oligonucleotide could be used in the assay (as a substrate) or purification (as an affinity probe) of the enzyme(s) responsible for eucaryotic mRNA splicing.

The authors wish to thank Dr. G. G. Brownlee for helpful discussions at the initiation of this project.

This work was supported by National Institutes of Health grants HD08188 to B. O'Malley and HD07495 (Baylor Center for Population Research and Reproductive Biology), American Cancer Society grant BC-101 and a Robert A. Welch Foundation grant Q-611 to A. Means. J. Catterall and J. Stein are recipients of National Institutes of Health postdoctoral fellowships.

Received for publication 21 April 1980.

REFERENCES

1. Alberts, H., M. Rabs, and H. J. Gross. 1978. Isolation from rat liver and sequence of a RNA fragment containing 32 nucleotides from position 5 to 36 from the 3' end of ribosomal 18S RNA. Nucl. Acids Res. 5:425-434.
2. Baralle, F. E., and G. G. Brownlee. 1978. AUG is the only recognizable signal sequence in the 5' non-coding regions of eukaryotic mRNA. Nature (Lond.) 274:54-56.
3. Böhm, G., and D. Dobberstein. 1975. Transfer of proteins across membranes. II. Reconstitution of functional rough microsomes from heterologous components. J. Cell Biol. 67:852-857.
4. Bolivar, F., R. L. Rodriguez, P. O. Green, M. D. Betlack, H. L. Heynecker, H. W. Boyer, J. H. Crosa, and S. Falkow. 1978. Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gen (Amst). 2:95-113.
5. Botchan, M., W. Topp, and J. Sambrook. 1976. The arrangement of Simian virus 40 sequences in the DNA of transformed cells. Cell. 9:269-287.
6. Brenchley, R., C. Benoit, K. O'Hare, J. Glanlon, and P. Chaiken. 1978. Ovalbumin
gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc. Natl. Acad. Sci. USA 75:4853-4857.

7. Buell, G. N., M. P. Wickers, J. Carbon, and R. T. Schimke. 1979. Isolation of recombinant plasmids bearing cDNA to hen ovomucoid and lysozyme mRNAs. J. Biol. Chem. 254: 9277-9280.

8. Catterall, J. F., B. W. O'Malley, M. A. Robertson, R. Staden, Y. Tanaka, and G. G. Brownlie. 1978. Nucleotide sequence homology at 12 intron-exon junctions in the chick ovalbumin gene. Nature (Lond.). 275:310-313.

9. Catterall, J. F., J. P. Stein, E. C. Lai, S. L. C. Woo, A. Dugaiczyk, M. L. Mace, A. R. Means, and B. W. O'Malley. 1979. The chick ovomucoid gene contains at least six intervening sequences. Nature (Lond.). 278:323-327.

10. Denhardt, D. 1966. A membrane filter technique for the detection of complementary DNA. Biochem. Biophys. Res. Commun. 23:641-646.

11. Efstratiadis, A., F. C. Kafatos, and T. Maniatis. 1977. The primary structure of rabbit beta-globin mRNA as determined from cloned DNA. Cell 10:571-585.

12. Fitch, N. M. 1976. Is there selection against wobble in codon-anticodon pairing? Nature (Lond.). 257:510-513.

13. Grunstein, M., and D. S. Hogness. 1975. Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proc. Natl. Acad. Sci. U.S.A. 72:3961-3965.

14. Heindell, H. C., A. Lin, G. V. Paddock, G. M. Studnicka, and W. A. Salser. 1978. The primary sequence of rabbit alpha-globin, mRNA. Cell 15:43-54.

15. Humphries, P., M. Cochet, A. Krust, P. Gerlinger, P. Kourilsky, and P. Chambon. 1977. Molecular cloning of extensive sequences of the in vitro synthesized chicken ovalbumin structural gene. Nucl. Acids Res. 4:2389-2406.

16. Hynes, N. E., B. Groner, A. E. Sippel, M. C. Nguyen-Huu, and G. Schultz. 1977. mRNA complexity and egg white protein mRNA content in mature and hormone-withdrawn oviduct. Cell 11:923-932.

17. Kato, L. W. J. Kohr, and M. Laskowski, Jr. 1978. Evolution of avian ovomucoids. FEBS (Fed. Eur. Biochem. Soc.) Proc. Meet. 87:197-206.

18. Katz, L. P., R. W. Williams, S. Sato, R. W. Lewiat, and D. R. Helinski. 1977. Purification and characterization of covalently closed replicative intermediates of Col E 1 DNA from Escherichia coli. Biochemistry. 16:1677-1683.

19. Konkel, D. A., S. M. Tighman, and P. Leder. 1978. The sequence of the chromosomal mouse beta-globin major gene: homologies in capping, splicing, and poly A sites. Cell 15:1125-1132.

20. Kozak, M. 1978. How do eucaryotic ribosomes select initiation regions in messenger RNA? Cell 15:1109-1123.

21. Lai, E. C., J. P. Stein, J. F. Catterall, S. L. C. Woo, M. L. Mace, A. R. Means, and B. W. O'Malley. 1979. Molecular structure and flanking nucleotide sequences of the natural chicken ovomucoid gene. Cell 18:829-842.

22. Maxam, A. M., and W. Gilbert. 1977. A new method for DNA sequencing. Proc. Natl. Acad. Sci. U.S.A. 74:560-564.

23. McReynolds, L. A., B. W. O'Malley, A. D. Nisbett, J. E. Forbergii, S. Fields, D. Givel, M. Robertson, and G. G. Brownlie. 1978. Sequence of chicken ovalbumin mRNA. Nature (Lond.). 273:723-728.

24. McReynolds, L. A., J. F. Catterall, and B. W. O'Malley. 1977. The ovalbumin gene: cloning of a complete ds cDNA in a bacterial plasmid. Gene (Amst.). 2:217-231.

25. Miller, J. R., E. M. Cartwright, G. G. Brownlee, N. F. Federoff, and D. D. Brown. 1976. The nucleotide sequence of soybean 5S DNA in Xestospongia testudinaria. J. Mol. Biol. 106:37-73.

26. Palmiter, R. D. 1972. Regulation of protein synthesis in chick oviduct. I. Independent regulation of ovomucin, conalbumin, ovomucoid, and lysozyme induction. J. Biol. Chem. 247:6450-6461.

27. Proudfoot, N. J., and G. G. Brownlee. 1976. 3' non-coding region sequences in eukaryotic messenger RNA. Nature (Lond.). 263:211-214.

28. Sanger, F., and A. R. Coulson. 1978. The use of thin acrylamide gels for DNA sequencing. FEBS (Fed. Eur. Biochem. Soc.) Lett. 87:107-110.

29. Schroeder, H. W., C. D. Liarakos, R. C. Gupta, K. Randerath, and B. W. O'Malley. 1979. The ribosome binding site of ovalbumin messenger RNA. Biochemistry. 18:5798-5808.

30. Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503-518.

31. Staden, R. 1977. Sequence data handling by computer. Nucl. Acids Res. 4:4037-4051.

32. Staden, R. 1978. Further procedures for sequence analysis by computer. Nucl. Acids Res. 5:1013-1015.

33. Stein, J. P., J. F. Catterall, S. L. C. Woo, A. R. Means, and B. W. O'Malley. 1978. Molecular cloning of ovomucoid gene sequences from partially purified ovomucoid mRNA. Biochemistry. 17:5763-5772.

34. Theobold, S. N., R. D. Palmiter, and K. A. Walsh. 1978. Precursor of egg white ovomucoid. J. Biol. Chem. 253:9018-9022.

35. Tongawa, S. A. M. Maxam, R. Tizard, O. Bernard, and W. Gilbert. 1978. Sequence of a mouse germ-line gene for a variable region of immunoglobulin light chain. Proc. Natl. Acad. Sci. U.S.A. 75:1485-1490.

36. Woo, S. L. C., T. Chandra, A. R. Means, and B. W. O'Malley. 1977. The ovomucoid gene: purification of the coding strand. Biochemistry. 16:5670-5676.