Laskey, R. A., & Earnshaw, W. C. (1980) *Nature (London) 286*, 763–767.
Laskey, R. A., Mills, A. D., & Morris, N. R. (1977) *Cell (Cambridge, Mass.) 10*, 237–243.
Leffak, M., Grainger, R., & Weintraub, H. (1977) *Cell (Cambridge, Mass.) 12*, 837–845.
Mathis, D., Oudet, P., & Chambon, P. (1980) *Prog. Nucleic Acid Res. Mol. Biol. 24*, 1–55.
McKnight, S., & Miller, O. (1977) *Cell (Cambridge, Mass.) 12*, 795–804.
Oliver, D., Granner, D., & Chalkley, R. (1974) *Biochemistry 13*, 746–749.
Panyim, S., & Chalkley, R. (1969) *Arch. Biochem. Biophys. 130*, 337–346.
Ruiz-Carrillo, A., Wangh, L. J., & Allfrey, V. G. (1975) *Science (Washington, D.C.) 190*, 117–128.
Ruiz-Carrillo, A., Jorcano, J. L., Eder, G., & Lurtz, R. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 3284–3288.
Seale, R. L. (1974) *Nature (London) 255*, 247–249.
Seale, R. L. (1976) *Cell (Cambridge, Mass.) 9*, 423–429.
Seale, R. L. (1978a) *Proc. Natl. Acad. Sci. U.S.A. 75*, 2717–2721.
Seale, R. L. (1978b) *Cell Nucl. 4*, 155–172.
Seale, R. L. (1981) *Nucleic Acids Res. 9*, 3151–3158.
Seale, R. L., & Simpson, R. T. (1975) *J. Mol. Biol. 94*, 479–501.
Seidman, M. M., Levine, A. J., & Weintraub, H. (1979) *Cell (Cambridge, Mass.) 18*, 439–449.
Senshu, T., Fukada, M., & Ohashi, M. (1978) *J. Biochem. (Tokyo) 84*, 985–988.
Shih, R. J., Smith, L. D., & Keem, K. (1980) *Dev. Biol. 75*, 329–342.
Stein, A. (1979) *J. Mol. Biol. 130*, 103–134.
Stein, A., Whitlock, J. P., Jr., & Bina, M. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 5000–5004.
Thomas, J. O., & Kornberg, R. D. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 2626–2630.
Todd, R. D., & Garrard, W. T. (1979) *J. Biol. Chem. 254*, 3074–3083.
Tsanev, G., Vassilev, L., & Russev, R. (1980) *Nature (London) 285*, 584–486.
Weintraub, H. (1976) *Cell (Cambridge, Mass.) 9*, 419–422.
Weintraub, H. (1979) *Nucleic Acids Res. 7*, 781–792.
Weintraub, H., & Van Lente, F. (1974) *Proc. Natl. Acad. Sci. U.S.A. 71*, 4249–4253.
Weisbrod, S., & Weintraub, H. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 630–634.
Woodland, H. R., & Adamson, E. D. (1977) *Dev. Biol. 57*, 118–135.
Worcel, A., Han, S., & Wong, M. L. (1978) *Cell (Cambridge, Mass.) 15*, 969–977.

# Complete Nucleotide Sequence of the Chicken Chromosomal Ovalbumin Gene and Its Biological Significance†

Savio L. C. Woo, Wanda G. Beattie, James F. Catterall, Achilles Dugaiczyk, Roger Staden, George G. Brownlee, and Bert W. O'Malley*

ABSTRACT: The nucleotide sequence of the entire chicken chromosomal ovalbumin gene has been determined. The gene is 7564 nucleotides in length to code for a mature messenger RNA of 1872 nucleotides. Comparison of the sequence at the 5′-terminal region of the gene with that reported by others has revealed multiple polymorphic nucleotides in the structural, intervening, and flanking DNA sequences. Some of the polymorphic sites occur at positions very close to splice junctions or the eucaryotic promoter sequence, yet apparently have little or no effect on the expression of this gene. The heptanucleotide promoter sequence TATATAT present in the 5′-flanking region of the ovalbumin gene does not occur within the confines of the gene. Nevertheless, multiple Hogness box sequences similar to those found in other eucaryotic genes were delineated within the boundaries of the gene. These internal Hogness box sequences are not used for transcription initiation. Similarly, the hexanucleotide sequence AATAAA common to all eucaryotic messenger RNAs at the 3′-untranslated region occurs seven additional times within the ovalbumin gene. These sites are not used for transcription termination or polyadenylation. Thus, although these sequences may play important roles in the initiation or termination of gene transcripts as well as polyadenylation of the transcripts, the specificity for such biological functions must not reside within these sequences alone. Furthermore, sequences complementary to the highly conserved rat U1 small nuclear RNA have been found throughout the gene. Many of these regions of complementarity occur in the structural sequences. If the small nuclear RNA does play a role in splicing, the specificity must be provided also by other as yet undefined components.

Expression of the ovalbumin gene in the chicken oviduct is regulated by steroid hormones (O'Malley & Means, 1974; Woo & O'Malley, 1975; Harris et al. 1973; Sullivan et al., 1973; Palmiter, 1975). Further insight into the molecular mechanism by which steroids regulate ovalbumin gene expression requires detailed knowledge with regard to the molecular structure and nucleotide sequence of the natural gene. We and others have previously reported the cloning and characterization of various overlapping genomic chick DNA fragments containing the ovalbumin gene, which had led to the discovery that the structural ovalbumin gene sequences are separated into eight segments by seven intervening DNA sequences (Woo et al., 1978; Dugaiczyk et al., 1978, 1979;

† From the Department of Cell Biology, Baylor College of Medicine, Houston, Texas 77030 (S.L.C.W., W.G.B., J.F.C., A.D., and B.W.O.), The Howard Hughes Medical Institute, Baylor College of Medicine, Houston, Texas 77030 (S.L.C.W.), and the MRC Laboratory of Molecular Biology, Hill Roads, Cambridge, United Kingdom (R.S. and G.G.B.). *Received April 8, 1981.* This work was supported by National Institutes of Health Grant HD08188 and the Baylor Center for Population Research and Reproductive Biology. S.L.C.W. is an Investigator of the Howard Hughes Medical Institute.

0006-2960/81/0420-6437$01.25/0 

Garapin et al., (1979); Mandel et al., 1978; Gannon et al., 1979). The intervening sequences, similar to the structural sequences, are unique chick DNA sequences which are transcribed in their entireties during gene expression and are inducible by steroid hormones (Woo et al., 1978). Subsquently we substantiated that the entire gene is transcribed into a large precursor molecule followed by excision of the intervening RNA sequences and appropriate ligation of the structural RNA sequences to produce the mature RNA (Roop et al., 1978; Tsai et al., 1980). Nucleotide sequencing analysis at the junctions between structural and intervening sequences within the gene has established that consensus sequences TCAGGT may be a signal for splicing and that RNA splicing occurs immediately 5′ to a GT doublet and 3′ to an AG doublet at the termini of an intervening sequence (Catterall et al. 1978; Breathnach et al., 1978).

In the present report, we define the entire sequence structure of the chromosomal ovalbumin gene and present comments on its polymorphic nature. Since it has been hypothesized that the small RNA U1 may promote splicing by forming a RNA/RNA duplex with the splice junctions of precursor RNAs through nucleotide complementarity (Lerner et al., 1980; Rogers & Wall, 1980), this hypothesis has been subjected to a search analysis of regions within the entire ovalbumin gene sequence that are capable of forming complementary duplexes with the rat U1 RNA sequence. In addition, the frequency of occurrence and significance of the hexanucleotide sequence AATAAA present at the 3′-untranslated region of all eucaryotic mRNAs (Proudfoot & Brownlee, 1976) and the "Hogness box" heptanucleotide sequence present at the 5′-flanking region of eucaryotic genes have been analyzed.

## Materials and Methods

*Materials.* Restriction enzymes were purchased from Bethesda Research Laboratories and New England Biolabs; DNase and bacterial alkaline phosphatase were from Worthington, and *Escherichia coli* DNA polymerase I, its Klenow fragment, and T4 polynucleotide kinase were from Boehringer Mannheim. Radioactive $^{32}$P-labeled deoxyribonucleoside triphosphates and [$^{32}$P]ATP were from Amersham. Xomat R X-ray film was from Kodak.

*Methods.* Plasmid DNA was prepared from *E.* Coli RR1 according to the method of Katz et al. (1973). Restriction mapping analytical and preparative gel electrophoreses were performed as reported previously (Lai et al., 1979a). End labeling and chemical degradation for DNA sequencing was according to the method of Maxam & Gilbert (1977), and thin polyacrylamide gel electrophoresis was by the method of Sanger & Coulson (1978).

## Results

*Structure and Sequencing of the Complete Gene.* The overall molecular structure of the ovalbumin gene, as determined previously by restriction mapping, electron microscopy, and limited nucleotide sequencing of the various cloned genomic fragments of the gene, is shown in the upper panel of Figure 1. Six independent clones of overlapping restriction fragments were used to construct the map, and the order of these clones from 5′ to 3′ of the genes is OV4.5, OV3.2, OV2.4, OV1.8, OV2.7, and OV4.8 (Woo et al., 1978; Dugaiczyk et al., 1978, 1979). The strategy of sequencing each of these DNA fragments is shown in the lower panels of Figure 1. OV2.4 and OV1.8 were the first fragments cloned, and their nucleotide sequences were determined in their entireties (Figure 1, panels 3 and 4). The nucleotide sequence of 1.2 kilobases (kb) of DNA 5′ to OV2.4 DNA was obtained from OV3.2 and OV4.5 DNA (Figure 1, panel 2). The 5′-terminal *Hind*III site in OV3.2 DNA was crossed by sequencing *Hph*I fragment generated from OV4.5 DNA. The 5′-terminal segment of the structural gene was located at 226 base pairs (bp) from the *Hind*III site and was uninterrupted by intervening sequences. The nucleotide sequence at the 3′ portion of the gene was determined mainly from OV2.7 and OV4.8 DNAs (Figure 1, panel 5). The 3′ terminus of the structural sequence was found within an *Hph*I fragment, thereby defining the overall domain of the ovalbumin gene.

Both *Eco*RI sites confining the OV1.8 DNA were sequenced across by labeling the neighboring *Alu*I sites located on the 3′ side of the corresponding *Eco*RI sites in OV4.8 DNA and sequencing the complementary strands (Figure 1, panel 5). Sequencing across the major restriction sites that were used for cloning was necessary in order to eliminate the possibility of the presence of a second restriction site so that an internal fragment could have been lost. The only such site not sequenced across was the *Eco*RI site at the 5′ terminus of OV2.4 DNA. Digestion of OV3.2 DNA with *Hph*I, however, has generated all of the fragments common with the corresponding region on OV2.4 DNA, those predicted from the rest of the OV3.2 DNA, and the expected 556-bp fragment containing that particular *Eco*RI site. Digestion of this fragment with *Eco*RI has generated only two bands at the proper sizes as analyzed by high percentage acrylamide gel electrophoresis.

The sequences of several regions internal to the structural gene segments were determined previously with a full-length cDNA clone (Catterall et al., 1978; McReynolds et al., 1978) and were not repeated with the genomic DNA clones. That there were no additional intervening DNA sequences within these regions was again established by restriction mapping. The only exception was a 13-bp *Hin*fI fragment that was present at the 3′-untranslated region (O'Hare et al., 1979). The presence of this nucleotide sequence has been confirmed by sequencing a fragment from OV2.7 DNA labeled at the *Xba*I site (Figure 1, panel 5).

*Nucleotide Sequence of the Entire Gene and Its Intrinsic Properties.* The nucleotide sequence of the entire ovalbumin gene is presented in Figure 2. The gene is 7564 bp in length to code for a messenger RNA of 1872 nucleotides. The regions in the gene corresponding to the mRNA sequences are underlined.

The exact size of the individual structural and intervening sequences and some other apparent properties of these individual regions are shown in Table I. The gene is comprised of 2465 adenines, 1450 cytosines, 1418 guanines, and 2251 thymines. The structural sequences are AT rich, with a range of 54% to 60%. The intervening sequences are even more AT rich (61–68%). Since the ovalbumin gene is comprised of 75.5% intervening sequences (6339 bp/7564 bp) and only 24.5% structural sequences (1872 bp/7564 bp), the entire gene is relatively rich in its AT content.

The distribution of dinucleotides within the gene is shown in Table II. The frequencies of occurrence of 15 out of 16 possible combinations do not deviate significantly from the expected values based on random distribution. The dinucleotide CG, however, occurs 42 times in the gene and constitutes only 0.6% of all the dinucleotides. This frequency is 17% of the expected random value. The low frequency for CG doublets is universal among eucaryotic genes and accounts for the lack of restriction cleavage sites for enzymes such as *Hha*I, *Hpa*II, and *Sst*II which have recognition sequences of GCGC, CCGG, and CCGCGG, respectively.


FIGURE 1: Strategy for sequencing of the gene. The structural and intervening sequences of the gene are represented by closed and open boxes, and their numbers are shown with the roman numerals I–VIII and the letters A–G, respectively. The sites of labeling are indicated by the vertical lines, and the corresponding restriction enzymes used were shown above the vertical lines. The arrows indicate the direction and extent of sequencing using those labeled DNA fragments. Labeling was carried out with [$\gamma$-$^{32}$P]ATP and T4 polynucleotide kinase; 3′ labeling was achieved by filling in the 3′-recessed DNA termini with appropriate $^{32}$P-labeled deoxynucleotides and the Klenow fragment of *E. coli* DNA polymerase I. For assurance of the accuracy of the sequences, a major portion of the gene was sequenced on both strands, as shown by overlapping arrows of opposite directions; other regions of the gene were sequenced on both strands by doing 5′ and 3′ labeling at the same restriction site. The sequence of the 5′ portion of the gene was derived from OV4.5 and OV3.2 DNAs (panel 2) and the 3′ portion from OV4.8 and OV2.7 DNAs (panel 5). The sequence of the middle portion of the gene was derived from OV2.4 and OV1.8 DNAs (panels 3 and 4). The strategy for sequencing OV2.4 DNA has been reported previously (Robertson et al., 1979) and is not shown in here. The sequence data were kept in a computer file and edited by using the SEQEDT program of Staden (Staden, 1977). Appropriate programs were used to search for restriction cleavage sites and other intrinsic properties of the gene.

Table I: Intrinsic Properties of the Ovalbumin Gene

| structural gene segment | intervening sequence | location in gene | exact size (nucleotides) | AT content (%) |
|---|---|---|---|---|
| I | | 1–47 | 47 | 58 |
| | A | 48–1636 | 1589 | 64 |
| II | | 1637–1821 | 185 | 54 |
| | B | 1822–2072 | 251 | 62 |
| III | | 2073–2123 | 51 | 56 |
| | C | 2124–2074 | 581 | 61 |
| IV | | 2705–2833 | 129 | 57 |
| | D | 2834–3233 | 400 | 68 |
| V | | 3234–3351 | 118 | 56 |
| | E | 3352–4309 | 958 | 61 |
| VI | | 4310–4452 | 143 | 57 |
| | F | 4453–4783 | 331 | 64 |
| VII | | 4784–4939 | 156 | 54 |
| | G | 4940–6521 | 1582 | 64 |
| VIII | | 6522–7564 | 1043 | 60 |

Table II: Distribution of Dinucleotides

| dinucleotide | no. | % | expected % |
|---|---|---|---|
| AA | 867 | 11.5 | 10.6 |
| AC | 408 | 5.4 | 6.2 |
| AG | 565 | 7.5 | 6.1 |
| AT | 625 | 8.3 | 9.7 |
| CA | 585 | 7.7 | 6.2 |
| CC | 247 | 3.3 | 3.6 |
| CG | 42 | 0.6 | 3.5 |
| CT | 555 | 7.3 | 5.6 |
| GA | 458 | 6.1 | 6.1 |
| GC | 320 | 4.2 | 3.5 |
| GG | 256 | 3.4 | 3.5 |
| GT | 384 | 5.1 | 5.6 |
| TA | 554 | 7.3 | 9.7 |
| TC | 455 | 6.0 | 5.6 |
| TG | 555 | 7.3 | 5.6 |
| TT | 687 | 9.1 | 8.9 |

Cleavage sites on the gene for commercially available restriction enzymes with known recognition sequences were determined from the gene sequence with the aid of a computer and are shown in Table III. Enzymes that do not cut the gene include *Bam*HI, which has proven useful in cloning of a 3.5-kb *Bam*HI Herpes virus DNA fragment containing the thymidine

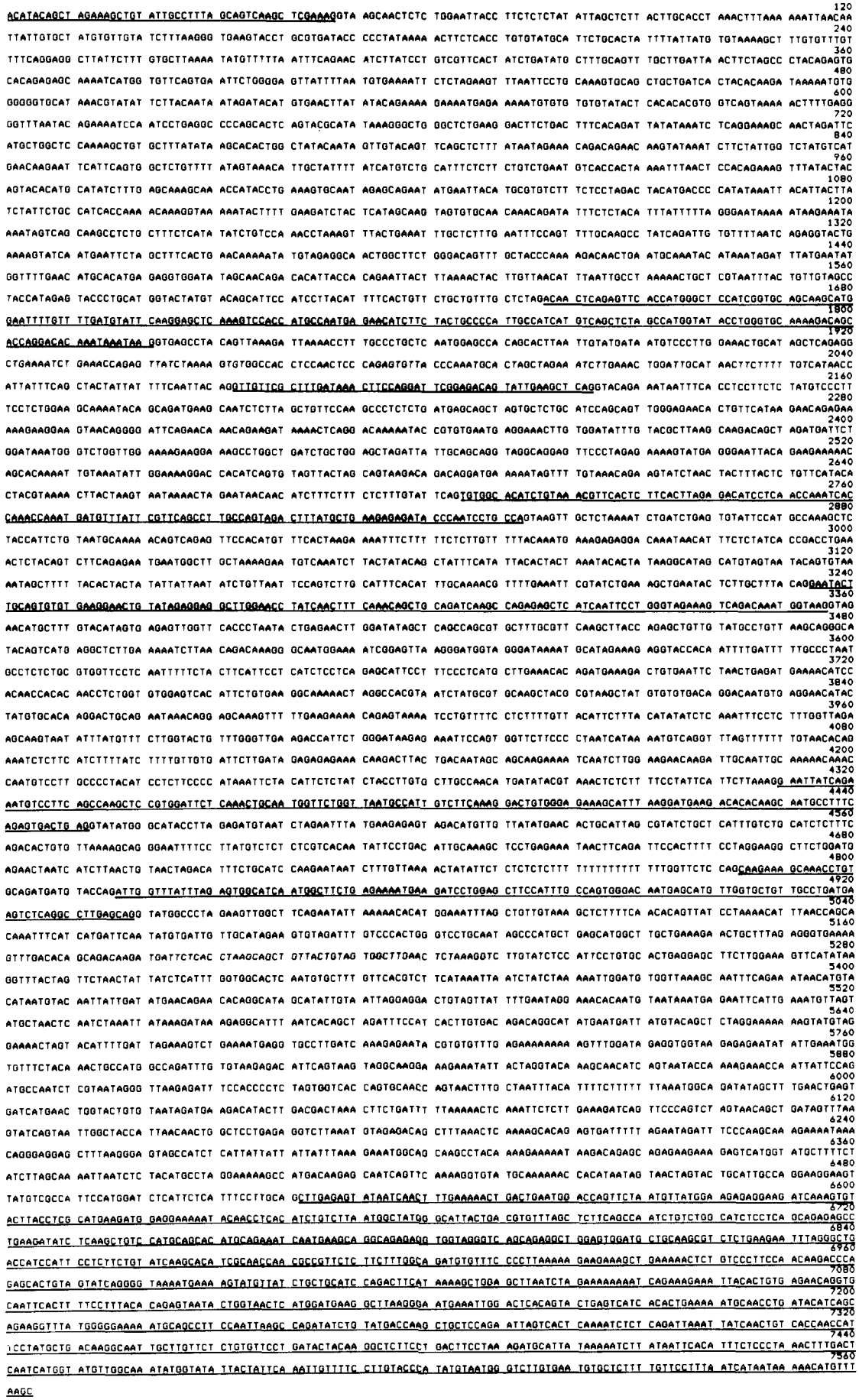

FIGURE 2: Primary sequence of the chromosomal ovalbumin gene. Nucleotide 1 is defined as the first nucleotide present in the mature ovalbumin mRNA as determined by McReynolds et al., (1978).

Table III: Positions of Restriction Enzyme Cleavage Sites in the Chromosomal Ovalbumin Gene[a]

| enzyme | nucleotide no.[b] |
|---|---|
| *Acc*I | 564, 1777 |
| *Alu*I | 7, 15, 38, 84, 227, 449, 735, 1340, 1706, 1764, 1911, 1993, 2049, 2117, 2200, 2227, 2387, 2451, 2876, 3059, 3124, 3211, 3296, 3316, 3417, 3443, 3453, 3794, 3805, 4336, 4628, 4869, 4989, 5000, 5197, 5567, 5617, 5986, 6107, 6179, 6249, 6520, 6678, 6734, 6927, 7024, 7030, 7259 |
| *Ava*II | 2547, 5100, 6559 |
| *Bal*I | 1954 |
| *Bcl*I | 455, 5686, 6000 |
| *Bgl*II | 1123 |
| *Bst*EII | 5925 |
| *Dde*I | 624, 638, 700, 1641, 1913, 2119, 2197, 2325, 2655, 2856, 2925, 3401, 3419, 3647, 3704, 4448, 4467, 4633, 4848, 4924, 5119, 5191, 5252, 5995, 6155, 6363, 6707, 7289 |
| *Eco*RI | 389, 847, 1333, 3695, 5501 |
| *Eco*RII | 1782, 1802, 2094, 2434, 3328, 4864 |
| *Fnu*4HI | 1669, 2225, 2463, 3297, 4149, 5195, 6297, 6518, 6744, 6810, 7003, 7223, 7260 |
| *Hae*III | 628, 1955, 3772, 4928, 4944, 5780 |
| *Hga*I | 6817 |
| *Hgi*AI | 358, 1705, 2232, 3315, 3844, 5247, 6961, 7522 |
| *Hin*cII | 1513 |
| *Hin*dIII | 226, 3442 |
| *Hin*fI | 716, 2098, 2300, 2395, 3744, 4110, 4345, 4649, 5054, 5182, 6341, 7150, 7163, 7437 |
| *Hpa*I | 1513 |
| *Hph*I | 149, 186, 922, 1093, 1649, 1821, 2138, 2757, 2988, 3390, 5153, 5187, 5927, 7310 |
| *Kpn*I | 3572 |
| *Mbo*II | 820, 907, 1121, 1736, 2315, 2511, 2729, 2810, 3010, 3883, 3999, 4034, 4086, 4170, 4223, 4382, 4417, 4492, 4858, 5338, 6029, 6334, 6579, 6587, 6614, 6681, 6722, 6825, 6853, 6889, 6921, 7374 |
| *Mnl*I | 247, 597, 626, 1215, 1312, 1365, 1461, 1917, 1960, 2141, 2162, 2213, 2350, 2499, 2746, 2966, 3265, 3268, 3490, 3570, 3602, 3617, 3638, 3646, 3664, 3733, 3830, 3910, 3947, 4221, 4450, 5150, 5254, 5456, 5552, 5677, 5731, 5917, 6159, 6245, 6584, 6606, 6621, 6706, 6777, 6795, 6851 |
| *Pst*I | 3298, 3855 |
| *Pvu*II | 448, 3295, 5196, 6106 |
| *Rsa*I | 155, 642, 774, 962, 1315, 1570, 1582, 1589, 2370, 3371, 3573, 3985, 4810, 5398, 5407, 5613, 5649, 5836, 6013, 6457, 6599, 7494 |
| *Sau*3A | 456, 1124, 2441, 2853, 3303, 4861, 5687, 6001, 6085, 6498, 6590 |
| *Sau*96I | 628, 2547, 4944, 6559 |
| *Sst*I | 1705, 3315 |
| *Taq*I | 41 |
| *Tha*I | 3799 |
| *Xba*I | 422, 1632, 4480, 7036 |

[a] No Cleavage Sites for *Ava*I, *Bam*HI, *Bgl*I, *Cla*I, *Hae*II, *Hha*I, *Hpa*II, *Mst*I, *Sal*I, *Sma*I, *Sst*II, *Xho*I, *Xor*II. [b] The number given is the first 5′ nucleotide of the recognition sequence and not the point of cleavage.

kinase gene and the entire ovalbumin gene in pBR322 DNA for transformation of mouse LTK cells (Lai et al., 1980). Other enzymes that cut the gene only once are *Taq*I (41), *Bgl*II (1123), *Hin*cII or *Hpa*I (1513), *Bal*I (1954), *Kpn*I (3572), *Tha*I (3799), *Bst*EII (5924), and *Hga*I (6817). Of these, the *Taq*I (41) site is of particular significance since it is located within the first structural gene segment and had been used as a diagnostic site to identify the 5′ end of the gene. In addition, this site is polymorphic since it is not present in the genes cloned by another laboratory.

*Polymorphism within the Gene. (1) Structural Sequence Polymorphism.* We have previously established a complete restriction map of the structural ovalbumin gene from its nucleotide sequence (McReynolds et al., 1978). The *Taq*I site at nucleotide 41 is present in the cDNA clones as well as the corresponding genomic clone (Dugaiczyk et al., 1979) but is absent in the structural ovalbumin gene cloned by O'Hare et al. (1979), and the chromosomal gene cloned by Gannon et al. (1979). A comparison of the nucleotide sequences between these clones has revealed a G–A transition at nucleotide 43 (Figure 3), resulting in the polymorphic *Taq*I site (TCGA) at this position. Since this polymorphism occurs within the 5′-untranslated region of the gene, there should be no change in the amino acid sequence of the protein and is therefore a silent mutation. Comparison of additional nucleotide sequences between the structural regions of the two clones has revealed another substitution at position 34, which is a G–C transversion (Figure 3). Again this polymorphism occurs in

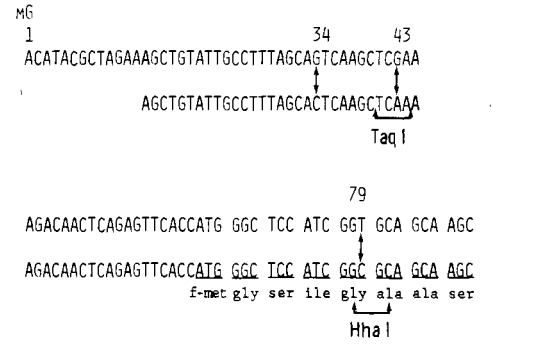


FIGURE 3: Exonic polymorphism. The 5′-structural sequences of the gene cloned by us (Catterall et al., 1977; McReynolds et al., 1978) and by O'Hare et al. (1979) are shown on the upper and lower lines, spectively. The polymorphic nucleotides are shown with double-headed arrows, and the nucleotide numbers shown on top of the polymorphic sites are according to McReynolds et al. (1978). The *Taq*I recognition sequence TCGA is present in the upper sequence, and the *Hha*I recognition sequence GCGC is present only in the lower sequence. The third polymorphic nucleotide at position 34 does not result in the creation or destruction of a restriction site.

the 5′-untranslated region of the gene and is silent. Also, the restriction enzyme *Hha*I does not cleave the structural ovalbumin gene (Catterall et al., 1978; McReynolds et al., 1978). An independent ovalbumin cDNA clone obtained by O'Hare et al. (1979), however, appears to contain a *Hha*I cleavage site (GCGC) at nucleotide 78–81. In this instance, a comparison of the two gene sequences has shown that there is a

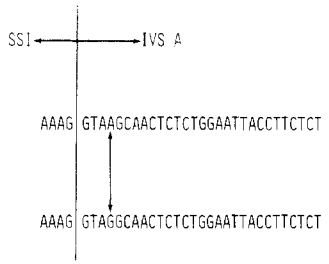


FIGURE 4: Intronic polymorphism. The upper and lower nucleotide sequences surrounding the junction between the first structural sequence (SSI) and first intervening sequence (IVS A) of the gene cloned by us (Dugaiczyk et al., 1979) and by Gannon et al. (1979), respectively. The alleged splice point in the gene is shown by the vertical line separating SSI and IVS A. The polymorphic nucleotide is shown with the double-headed arrow.

T–C transition at nucleotide 79 (Figure 3). The transition occurs at the third position of the codon for glycine and is an example of a classical silent mutation due to the degeneracy of the genetic code. There are two additional silent mutations in the structural gene, one being an A–G transition at the third base of the codon for glutamine (nucleotide 226) and the other being an A–C transversion in the 3′-untranslated regions of the gene. A sixth polymorphic nucleotide at position 626 (G–A transition), however, would have caused the change from an alanine codon GCA (McReynolds et al., 1978) to a threonine codon ACA (O'Hare et al., 1979). Barring a mistake in the nucleotide sequence, there should be two types of the ovalbumin polypeptide.

(*2*) *Intervening Sequence Polymorphism.* We have reported previously that polymorphism in the intervening sequences of the ovalbumin gene exists between individual animals of the same species (Lai et al., 1979b). This type of polymorphism creates genotypic alleles but does not result in any phenotypic manifestation. In addition to the A–G transition within intervening sequence E of the ovalbumin gene, which has resulted in the appearance of an additional *Eco*RI cleavage site, and some other nucleotide conversions in intervening sequence A, which have caused the appearance of polymorphic *Hae*III sites (Lai et al., 1979b), other intronic polymorphism has been identified. A *Mbo*II site is present at about 50 bp to the left of the unique *Hind*III site in intervening sequence A of the ovalbumin gene cloned by Gannon et al. (1979) but is absent in the gene cloned by us (Figure 4). Furthermore, there is yet another A–G transition at nucleotide 4 in the intervening sequence A. The existence of at least two polymorphic sites within a stretch of about 120 nucleotides in this particular intervening sequence of the two independently cloned chromosomal ovalbumin genes confirms our previous notion that the frequency of this type of polymorphism must be rather high within intervening sequences.

(*3*) *Flanking Sequence Polymorphism.* The nucleotide sequences at the 5′-flanking region of the ovalbumin genes cloned by us and Gannon et al. (1979) are shown in Figure 5, and polymorphism in this region is again common. In addition to a G–C transversion at nucleotide −61, there is a G–A transition at nucleotide −34. This transition is only two nucleotides from the Hogness box sequence TATATAT which is located at positions −32 to −26 from the 5′ end of the gene (Figure 5). It has been suggested that the GC-rich regions surrounding the *E. coli lac* promoter sequence may be an important feature for RNA polymerase binding (Dickson et al., 1975). This polymorphic nucleotide, however, indicates that the most critical domain within the gene for RNA polymerase recognition and transcription initiation may not exceed two nucleotides beyond the Hogness box.

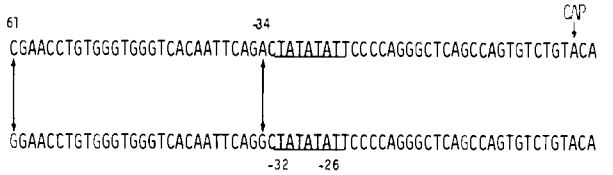


FIGURE 5: Flanking sequence polymorphism. The upper 5′-flanking sequence was obtained from pOV4.5 cloned by Dugaiczyk et al. (1979), and the lower sequence was from Gannon et al. (1979). The start site for the structural sequence is indicated by the CAP arrow. Polymorphic nucleotides are shown with double-headed arrows, and the Hogness box sequence TATATAT are underlined.

*Hogness Box Sequences within the Ovalbumin Gene.* At 32 bp preceding the 5′ terminus of the ovalbumin gene, the heptanucleotide sequence TATATAT occurs. This heptanucleotide resembles the Hogness box sequence common among eucaryotic genes transcribed by RNA polymerase II. Owing to its similarity with the Pribnow box sequence which is a recognized procaryotic promoter for transcription (Pribnow, 1975), a eucaryotic promoter function of the Hogness box sequence has been implicated. Furthermore, recent cell-free transcription experiments of cloned eucaryotic gene fragments have suggested that the Hogness box sequence is indeed required for proper RNA initiation (Weil et al., 1979; Manley et al., 1980; Luse & Roeder, 1980; Wasylyk et al., 1980; Talkington et al., 1980; Tsai et al., 1981). We have attempted to determine indirectly whether the Hogness box sequence alone is sufficient for correct transcription initiation by searching for known Hogness box sequences within the chromosomal ovalbumin gene (Table IV).

Although there is no TATATAT sequence within the ovalbumin gene, such a sequence occurs about 1000 bp further upstream from the 5′ end of the gene. In addition, a search for 11 additional established Hogness box sequences has identified a total of 17 such sequences internal to the gene, of which 13 are present in the various intervening sequences and 4 in structural sequence VIII (Table IV). It is also interesting to note that all four Hogness box sequences present in structural sequence VIII are limited to the 3′-untranslated region of the gene. The possibility of these internal promoter sequences not being utilized for lack of an initiating purine 30 nucleotides away appears unlikely since a search for purines at those positions has revaled ample AC and AT dinucleotides which are known initiation nucleotides for many eucaryotic genes (Table V). These observations would argue that the Hogness box sequence alone cannot be a sufficient and universal signal for initiation of transcription.

*AATAAA Hexanucleotide within the Ovalbumin Gene.* The obligatory presence of the hexanucleotide AAUAAA in the 3′-untranslated region of eucaryotic mRNAs has led to the hypothesis that it is a signal for either transcription termination or polyadenylation (Proudfoot & Brownlee, 1976). The fact that sea urchin histone genes H2A and H3 do not possess the hexanucleotide sequence and are not polyadenylated would lend some support to the hypothesis, although the H2B gene does contain such a sequence in the 3′-untranslated region at 97 nucleotides following the termination codon TAG (Sures et al., 1978). Whether the hexanucleotide alone is sufficient for signaling such biological processes has been analyzed indirectly by searching for such sequences within the ovalbumin gene. In addition to the AATAAA hexanucleotide present at the 3′-untranslated region of the structural gene, there are seven more such sequences scattered among intervening and structural sequences (Table VI). Since aborted RNA transcripts of the ovalbumin gene have not been detected in the oviduct cell, the presence of

Table IV: Eucaryotic Promoter Sequences within the Chromosomal Ovalbumin Gene

| eucaryotic genes | promoter sequence[a] | present at position of ovalbumin gene | location in structural sequence | location in intervening sequence |
|---|---|---|---|---|
| chick ovalbumin, chick ovomucoid | TATATAT | 0 | | |
| Ad 2 major late chick conalbumin *B. mori* fibroin | TATAAAA | 174 | | A |
| | | 7399 | VIII | |
| chick insulin | TATAATT | 7410 | VIII | |
| chick β-globin | GATAAAA | 469 | | A |
| | | 2318 | | C |
| | | 3553 | | E |
| rabbit β-globin | CATAAAA | 4046 | | E |
| | | 7018 | VIII | |
| mouse β-globin, major | CATATAA | 647 | | A |
| | | 1061 | | A |
| | | 5274 | | G |
| mouse β-globin, minor | TATATAA | 691 | | A |
| | | 745 | | A |
| mouse immunoglobulin V λII | TATATTA | 78 | | A |
| | | 3139 | | D |
| | | 7467 | VIII | |
| rat insulin | TATAAAG | 5540 | | G |
| sea urchin and *Drosophila* histones | TATAAGT | 0 | | |
| | TATAAAC | 0 | | |
| | TATAAAT | 0 | | |

[a] These sequences were derived from the following studies: Gannon et al. (1979); Lai et al. (1979a); Ziff & Evans (1978); Bernard et al. (1978); Konkel et al. (1978); Hardison et al. (1979); Van den Berg et al. (1978); Day et al. (1981); Lomedico et al. (1979); Cordell et al. (1979); Perler et al. (1980); Cochet et al. (1979); Tsujimoto & Suzuki (1979); Schaffner et al. (1978); Sures et al. (1978).

Table V: Nucleotide Sequences Neighboring the Eucaryotic Promoters

| position in gene | nucleotide sequence (1 11 21 31) |
|---|---|
| 78 | ACCTTCTCTCTATATTAGCTCTTACTTGCACCTAAACTTTAAAAAATT |
| 174 | TGATACCCCCTATAAAAACTTCTCACCTGTGTATGCATTCTGCACTAT |
| 469 | CACTACACAAGATAAAAATGTGGGGGGTGCATAAACGTATATTCTTAC |
| 649 | TCAGTACGCATATAAAGGGCTGGGCTCTGAAGGACTTCTGACTTTCAC |
| 691 | TTTCACAGATTATATAAATCTCAGGAAAGCAACTAGATTCATGCTGGC |
| 745 | AGCTGTGCTTTATATAAGCACACTGGCTATACAATAGTTGTACAGTTC |
| 1061 | TACATGACCCCATATAAATTACATTACTTATCTATTCTGCCATCACCA |
| 2318 | ACAAACAGAAGATAAAACTCAGGACAAAAATACCGTGTGAATGAGGAA |
| 3139 | TTTACACTACTATATTATTAATATCTGTTAATTCCAGTCTTGCATTTC |
| 3553 | GGATGGTAGGGATAAAATGCATAGAAAGAGGTACCACAATTTTGATTT |
| 4046 | TTCCCCTAATCATAAAATGTCAGGTTTAGTTTTTTTGTAACACAGAAA |
| 5274 | TTGGAAAGTTCATATAAGGTTTACTAGTTCTAACTATTATCTCATTTG |
| 5540 | CAATCTAAATTATAAAGATAAAGAGGCATTTAATCACAGCTAGATTTC |
| 7018 | ATCCAGACTTCATAAAAGCTGGAGCTTAATCTAGAAAAAAAATCAGAA |
| 7399 | AAAGATGCATTATAAAAATCTTATAATTCACATTTCTCCCTAAACTTT |
| 7410 | ATAAAAATCTTATAATTCACATTTCTCCCTAAACTTTGACTCAATCAT |
| 7467 | GCAAATATGGTATATTACTATTCAAATTGTTTTCCTTGTACCCATATG |

multiple AATAAA hexanucleotides within the gene would suggest that the seven additional hexanucleotides are not being used or recognized for termination of transcription or polyadenylation. Therefore, it appears that the hexanucleotide alone is not sufficient for specifying such biological functions.

*Potential Splice Junctions in the Ovalbumin Gene.* Due to the complementarity between the rat U1 small nuclear RNA to the intervening sequences at splice junctions of eucaryotic genes, it has been postulated that this small RNA molecule may provide the specificity for splicing of eucaryotic precursor RNAs to mature mRNAs. Consequently, the complementarity of rat U1 RNA with the intervening sequences at the 14 legitimate splice junctions of the ovalbumin gene was analyzed. When G–U base pairing is allowed for but looping out of nucleotides is disallowed, the complementarity between the 5′ octanucleotides of the seven intervening sequences with rat U1 RNA ranges from four to six nucleotides (Table VII). By use of the same percent complementarity, potential splice junctions with the first two nucleotides being GT in the ovalbumin gene were tabulated. At 50% complementarity level, all 7 legitimate splice junctions were

Table VI: Presence of the AATAAA Hexanucleotide within the Chromosomal Ovalbumin Gene

| nucleotide no. | location in gene: structural sequence | location in gene: intervening sequence |
|---|---|---|
| 1184 | | A |
| 1197 | | A |
| 1812 | II | |
| 2661 | | C |
| 3861 | | E |
| 5492 | | G |
| 6235 | | G |
| 7546 | VIII | |

detected among a total of 294 potential splice junctions, of which 68 are present in structural sequences. At 62.5% complementarity level, only five of the seven legitimate splice junctions were detected, although the number of total potential splice junctions has been reduced to 168, of which 28 are present in structural sequences. At 75% complementarity, there were only 61 potential splice junctions observable, but

Table VII: Potential Splice Junctions in the Chromosomal Ovalbumin Gene As Predicted from Complementarity with Rat U1 snRNA

| legitimate splice junctions | nucleotide sequence[a] | intervening sequence | complementarity with rat U1 snRNA: no. of nucleotides | complementarity with rat U1 snRNA: % | frequency of potential splice junction in gene[b]: IVS | SS | total | (legitimate splice junctions found)/(total legitimate junctions present) | (legitimate splice junctions found)/(total potential junction present) (%) |
|---|---|---|---|---|---|---|---|---|---|
| SS/IVS | GUACAGAA | C | | | | | | | |
| | GUAUGGCC | G | 4 | 50.0 | 226 | 68 | 294 | 7/7 | 2.4 |
| | GUAAGGUA | E | | | | | | | |
| | GUAUAUGG | F | 5 | 62.5 | 140 | 28 | 168 | 5/7 | 3.0 |
| | GUAAGCAA | A | | | | | | | |
| | GUGAGCCU | B | 6 | 75.0 | 53 | 8 | 61 | 3/7 | 4.9 |
| | GUAAGUUG | D | | | | | | | |
| IVS/SS | UGUUUGCUCUAG | A | | | | | | | |
| | UUUCAAUUACAG | B | | | | | | | |
| | CUUUGUAUUCAG | C | 8 | 66.6 | 70 | 21 | 91 | 7/7 | 7.7 |
| | CUUGCUUUACAG | D | | | | | | | |
| | UCAUUCUUAAAG | E | | | | | | | |
| | UUGGUUUCUCCAG | F | | | | | | | |
| | AUUUCCUUGCAG | G | 9 | 75.0 | 20 | 9 | 29 | 1/7 | 3.5 |

[a] Underlined nucleotides are complementary with rat U1 snRNA. [b] Number of regions in the gene capable of forming duplex with rat U1 snRNA with corresponding percentage of complementarity.

8 are still present in structural sequences, and 4 legitimate splice junctions have been missed. Thus, as the level of complementarity increases, there appears to be an inverse relationship between the ratio of (legitimate junctions found)/(legitimate junctions present) to the ratio of (legitimate junctions found)/(potential junctions present). Such a relationship is not apparent between the 3′-terminal dodecanucleotides and rat U1 RNA (Table VII). Six of the seven legitimate splice junctions contain eight nucleotides complementary to rat U1 RNA. At 66.6% complementarity, there are a total of 91 potential splice junctions with the last two nucleotides being AG; 21 of these are located in structural sequences (Table VII). At 75% complementarity, there are a total of 29 potential splice junctions, of which 9 are present in the structural sequences, even though the other 6 legitimate splice junctions have been missed. The detection of potential splice junctions within structural sequences at the highest level of stringency would suggest that complementarity with U1 RNA may not be a universal signal for RNA splicing.

## Discussion

By use of cloned overlapping DNA fragments of the ovalbumin gene, the nucleotide sequence of the entire chromosomal gene has been established. Since sequencing was performed on both strands over the major portion of the gene including those containing the polymorphic nucleotides, the accuracy of the sequence should be rather good. The chromosomal gene is 7564 bp in length and codes for a mature mRNA of 1872 nucleotides, with the remainder of the sequence being distributed among 7 intervening sequences ranging from 52 to 1589 nucleotides in length. The nucleotide content of the gene is 62.3% AT, reflecting the rather high AT content in the intervening sequences.

Polymorphism appears to be a rather common feature of the ovalbumin gene. Two polymorphic nucleotides are present between two independently cloned genes within the 5′-untranslated region (64 nucleotides) of the structural gene. Since both species of chicken produce large quantities of the ovalbumin polypeptide, the mutations at these particular locations presumably have little or no significant effect on ribosome binding or translation efficiency of the two messenger RNAs. We have recently isolated 87- and 92-nucleotide T1-resistant RNA fragments from ovalbumin mRNA after they were bound to the wheat germ 40S ribosomal subunit and its associated initiation factors (Schroeder et al., 1979). Both of these fragments contain the 5′ end of the mRNA through the initiation codon AUG. Structural and sequence analysis of these RNA fragments have indicated that there could be substantial base pairing between the 5′ end of the mRNA with the 3′ end of 18S rRNA. The base substitution at nucleotide 43 appears to have occurred in the region of the mRNA that has no complementarity with the 3′ end of 18S mRNA and should presumably not affect the efficiency of ribosome binding. The G–C transversion at nucleotide 34, however, would have decreased the 13-bp mRNA/rRNA region by 1 G/C pair. The effect of such a mutation on the stability of the entire mRNA/rRNA complex may not be sufficient to reduce the translational efficiency of the mRNA to a significant extent. It would, however, be interesting to compare the relative translational efficiency of the two ovalbumin mRNAs under a variety of conditions in order to investigate the significance of such a base substitution in the leader sequence of a eucaryotic gene.

In addition to the silent mutations in the structural gene, polymorphism within intervening sequences has proven to be a frequent event (Lai et al., 1979b). The significance of the

polymorphic nucleotide at position 4 in intervening sequence A, however, is beyond polymorphism per se. The fact that a nucleotide substitution can occur at a position so close to a splice point and yet presumably allow proper splicing could indicate that the absolute sequence signal for splicing may not extend very much farther than the GT dinucleotides at the 5′ end of an intervening sequence. On the other hand, since the polymorphic nucleotide at this critical position is an A–G transition and the corresponding nucleotide in U1 RNA is a uridine, it may not have a detrimental effect on splicing if the signal is indeed provided by base complementation with U1 snRNA and G–U base pairing exists in the cell.

There are two polymorphic nucleotides within the first 61 nucleotides at the 5′-flanking region. The A–G transition at nucleotide −34 is of particular significance, owing to its relationship with the Hogness box sequence at −32 to −26. This presumably silent mutation would again suggest that the most critical domain of the promoter may not extend by more than one or two nucleotides 5′ from the Hogness box sequence itself. Such a conclusion appears to be supported by the capacity for correct initiation in a cell-free transcription system of the cloned ovalbumin gene in which the nucleotide sequences 5′ to the Hogness box had been deleted (Tsai et al., 1981).

The Hogness box alone, however, does not appear to be the only signal required for transcription initiation because many such sequences occur throughout the ovalbumin gene and appear not be utilized in the intact oviduct cell. Furthermore, the failure of RNA polymerase to utilize these internal promoter sequences is not due to the lack of an initiating purine 30 nucleotides 3′ from the box. In fact, the Hogness box sequence of the major late adenovirus II gene TATAAAA is present at nucleotide 174–181 in the first intervening sequence of the gene. This promoter sequence not only is not recognized in vivo but also is not recognized in the cell-free transcription system (Tsai et al., 1981). Thus, whatever the additional signal(s) is (are) for transcription, it can be identified by a combination of site directed mutgenesis and gene transcription assays such as cell-free transcription or DNA-mediated gene transfer.

The presence of multiple hexanucleotide AATAAA sequence internal to the gene and the lack of evidence that they are recognized in the oviduct cell again suggest that its universal presence in the 3′-untranslated region of eucaryotic mRNAs may not be the only signal required for either transcription termination or polyadenylation. On the other hand, although the U1 small nuclear RNA has been postulated to play an important role in specifying RNA splicing through base complementation, the fact that many such nonfunctional complementary regions are present in the structural gene sequence would argue that U1 RNA alone cannot provide for all of the specificity for RNA splicing, if it is involved in splicing at all. From these analyses, a generalized conclusion may be reached: as important as all of these consensus sequences are, they cannot be the only signals specifying the alleged biological functions.

Acknowledgments

We thank Abigail Hirst, Serlina Robinson, and Susan Wildin for the excellent assistance in DNA sequencing and Sharon Moore for plasmid DNA preparation.

References

Bernard, O., Hozumi, N., & Tonegawa, S. (1978) *Cell (Cambridge, Mass.) 15*, 1133–1144.

Breathnach, R., Bendist, C., O'Hare, K., Gannon, F., & Chambon, P. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 4853–4857.

Catterall, J. F., O'Malley, B. W., Robertson, M. A., Staden, R., Tanaka, Y., & Brownlee, G. G. (1978) *Nature (London) 257*, 510–513.

Cochet, M., Gannon, F., Hen, R., Marotezux, L., Perrin, F., & Chambon, P. (1979) *Nature (London) 282*, 567–574.

Cordell, B., Bell, G., Tischer, E. DeNoto, F. M., Ullrich, A., Pictet, R., Rutter, N. J., & Goodman, H. M. (1979) *Cell (Cambridge, Mass.) 18*, 533–543.

Day, L., Hirst, A. J., Lai, E. C., Mace, M. L., & Woo, S. L. C. (1981) *Biochemistry 20*, 2091–2098.

Dickson, R. C., Abelson, J., Barnes, W., & Regnikoff, W. S. (1975) *Science (Washington, D.C.) 187*, 27–35.

Dugaiczyk, A., Woo, S. L. C., Lai, E. C., Mace, M. L., Jr., McReynolds, L., & O'Malley, B. W. (1978) *Nature (London) 274*, 328–333.

Dugaiczyk, A., Woo, S. L. C., Colbert, D. A., Lai, E. C., Mace, M. C., Jr., & O'Malley, B. W. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 2253–2257.

Gannon, F., O'Hare, K., Perrin, F., LePennec, J. P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B., & Chambon, P. (1979) *Nature (London) 278*, 428–434.

Garapin, A. C., Cami, B., Roskam, W., Kourilsky, P., LePennec, J. P., Perrin, F., Gerlinger, P., Cochet, M., & Chambon, P. (1978) *Cell (Cambridge, Mass.) 14*, 629–639.

Hardison, R. C., Butler, E. T., III, Lacy, E., Maniatis, T., Rosenthal, N., & Efstradiatis, A. (1979) *Cell (Cambridge, Mass.) 18*, 1285–1297.

Harris, S. E., Means, A. R., Mitchell, W. M., & O'Malley, B. W. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 3776–3780.

Katz, L., Kingsbury, D. T., & Helinski, D. R. (1973) *J. Bacteriol. 114*, 577–591.

Konkel, D. A., Tilghman, S. M., & Leder, P. (1978) *Cell (Cambridge, Mass.) 15*, 1125–1132.

Lai, E. C., Stein, J. P., Catterall, J. F., Woo, S. L. C., Mace, M. L., Means, A. R., & O'Malley, B. W. (1979a) *Cell (Cambridge, Mass.) 18*, 829–842.

Lai, E. C., Woo, S. L. C., Dugaiczyk, A., & O'Malley, B. W. (1979b) *Cell (Cambridge, Mass.) 16*, 201–211.

Lai, E. C., Woo, S. L. C., Bordelon-Riser, M. E., Fraser, T. H., & O'Malley, B. W. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 244–248.

Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L., & Steitz, J. A. (1980) *Nature (London) 283*, 220–224.

Lomedico, P., Rosenthal, N., Efstradiatis, A., Gilbert, W., Kolodner, R., & Tizard, R. (1979) *Cell (Cambridge, Mass.) 18*, 545–558.

Luse, D. S., & Roeder, R. G. (1980) *Cell (Cambridge, Mass.) 20*, 691–699.

Mandel, J. L., Breathnach, R., Gerlinger, P., LeMeur, M., Gannon, F., & Chambon, P. (1978) *Cell (Cambridge, Mass.) 14*, 641–653.

Manley, J. L., Fire, A., Cano, A., Sharp, P. A., & Gefter, M. L. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 3855–3859.

Maxam, A. M., & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 560–564.

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.

O'Hare, K., Breathnach, R., Benoist, C., & Chambon, P. (1979) *Nucleic Acids Res. 7*, 321–334.

O'Malley, B. W., & Means, A. R. (1974) *Science (Washington, D.C.) 182*, 610–620.

Palmiter, R. E. (1975) *Cell (Cambridge, Mass.)* *4*, 189-197.
Perler, F., Efstradiatis, A., Lomedico, P., Gilbert, W., Kolodner, R., & Dodgson, J. (1980) *Cell (Cambridge, Mass.)* *20*, 555-566.
Pribnow, D. (1975) *Proc. Natl. Acad. Sci. U.S.A.* *72*, 784-788.
Proudfoot, N. J., & Brownlee, G. G. (1976) *Nature (London)* *263*, 211-214.
Robertson, M. A., Staden, R., Tanaka, Y., Catterall, J. F., O'Malley, B. W., & Brownlee, G. G. (1979) *Nature (London)* *278*, 370-372.
Rogers, J., & Wall, R. (1980) *Proc. Natl. Acad. Sci. U.S.A.* *77*, 1877-1879.
Roop, D. R., Nordstrom, J. L., Tsai, S. Y., Tsai, M.-J., & O'Malley, B. W. (1978) *Cell (Cambridge, Mass.)* *15*, 671-685.
Sanger, F., & Coulson, A. R. (1978) *FEBS Lett.* *87*, 107-110.
Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H. O., & Birnstiel, M. L. (1978) *Cell (Cambridge, Mass.)* *14*, 655-671.
Schroeder, H. W., Liarakos, C. D., Gupta, R. C., Randerath, K., & O'Malley, B. W. (1979) *Biochemistry* *18*, 5798-5808.
Staden, R. (1977) *Nucleic Acids Res.* *4*, 4037-4051.
Sullivan, D., Palacios, R., Stavnezer, J., Taylor, J. M., Faras, A. J., Kiely, M. L., Summers, N. M., Bishop, J. M., & Schimke, R. T. (1973) *J. Biol. Chem.* *248*, 7530-7539.
Sures, I., Lowry, J., & Kedas, L. H. (1978) *Cell (Cambridge, Mass.)* *15*, 1033-1044.
Talkington, C. A., Nishioka, Y., & Leder, P. (1980) *Proc. Natl. Acad. Sci. U.S.A.* *77*, 7132-7136.
Tsai, M. J., Ting, A., Nordstrom, J., Zimmer, W., & O'Malley, B. W. (1980) *Cell (Cambridge, Mass.)* *22*, 219-230.
Tsai, S., Tsai, M.-J., & O'Malley, B. W. (1981) *Proc. Natl. Acad. Sci. U.S.A.* *78*, 879-883.
Tsujimoto, Y., & Suzuki, Y. (1979) *Cell (Cambridge, Mass.)* *16*, 425-436.
Van den Berg, J., Van Oogen, A., Mantei, N., Schambock, A., Grosveld, G., Flavell, R. A., & Weissmann, C. (1978) *Nature (London)* *276*, 37-44.
Wasylyk, B., Kedinger, C., Corden, J., Brison, O. and Chambon, P. (1980) *Nature (London)* *285*, 367-373.
Weil, P. A., Luse, D. S., Segall, J., & Roeder, R. G. (1979) *Cell (Cambridge, Mass.)* *18*, 469-484.
Woo, S. L. C., & O'Malley, B. W. (1975) *Life Sci.* *7*, 1039-1048.
Woo, S. L. C., Dugaiczyk, A., Tsai, M.-J., Lai, E. C., Catterall, J. F., & O'Malley, B. W. (1978) *Proc. Natl. Acad. Sci. U.S.A.* *75*, 6988-3692.
Ziff, E. B., & Evans, R. M. (1978) *Cell (Cambridge, Mass.)* *15*, 1463-1475.

# Structure of Ferric Pseudobactin, a Siderophore from a Plant Growth Promoting *Pseudomonas*†

Martin Teintze, M. B. Hossain, C. L. Barnes, John Leong,* and Dick van der Helm*

ABSTRACT: Both plant growth promoting *Pseudomonas* B10 and its yellow-green, fluorescent iron transport agent (siderophore) pseudobactin enhance the growth of the potato and control certain phytopathogenic microorganisms. The structure of the title compound has been determined by single-crystal X-ray diffraction methods using counter data. The structure consisted of a linear hexapeptide, L-Lys-D-*threo*-β-OH-Asp-L-Ala-D-*allo*-Thr-L-Ala-D-$N^{\delta}$-OH-Orn, in which the $N^{\delta}$-OH nitrogen of the ornithine was cyclized with the C-terminal carboxyl group, and the $N^{\epsilon}$-amino group of the lysine was linked via an amide bond to a fluorescent quinoline derivative. The iron-chelating groups consisted of a hydroxamate group derived from $N^{\delta}$-hydroxyornithine, an α-hydroxy acid derived from β-hydroxyaspartic acid, and an *o*-dihydroxy aromatic group derived from the quinoline moiety. The combination of metal-chelating ligands and the alternating L- and D-amino acids was unusual. The title compound crystallized as a single coordination isomer with the Λ absolute configuration. The present study is the first structural determination of a fluorescent siderophore. In the crystal structure, ferric pseudobactin formed a dimer, which constituted the asymmetric unit. The asymmetric unit also contained 26 water molecules. The molecules in the dimer were related by a pseudo-2-fold symmetry axis. Red-brown crystals of ferric pseudobactin ($C_{42}H_{57}N_{12}O_{16}Fe\cdot 13H_2O$), obtained from pyridine–acetic acid buffer solution equilibrated with water, conformed to space group *I*2 with $a$ = 29.006 (23) Å, $b$ = 14.511 (13) Å, $c$ = 28.791 (21) Å, and $\beta$ = 96.06 (5)° at −135 (2) °C. For eight molecules per unit cell, the calculated density was 1.38 g/cm³; the observed density was 1.40 g/cm³. The structure was refined by least-squares methods with anisotropic thermal parameters for all nonhydrogen atoms to a final $R$ factor of 0.08 (8989 observed reflections).

Specific strains of the *Pseudomonas fluorescens-putida* group have recently been used as seed inoculants on crop plants to promote growth and increase yields. These pseudomonads, termed plant growth promoting rhizobacteria (PGPR),[1] rapidly colonize plant roots of the potato, sugar beet, and radish and cause statistically significant yield increases (Kloepper et al.,

† From the Department of Chemistry, University of California at San Diego, La Jolla, California 92093 (M.T. and J.L.), and the Department of Chemistry, University of Oklahoma, Norman, Oklahoma 73019 (M.B.H., C.L.B., and D.v.d.H.). *Received March 24, 1981.* This work was supported in part by grants from the U.S. Public Health Service (AI 14084 and GM 21822) and National Institutes of Health (RR00719) to the Bioorganic, Biomedical Mass Spectrometry Resource (A. L. Burlingame, Director), Space Sciences Laboratory, University of California, Berkeley, CA 94720.

[1] Abbreviations used: PGPR, plant growth promoting rhizobacteria; PITC, phenyl isothiocyanate; $CF_3COOH$, trifluoroacetic acid; $N^{\delta}$-OH-Orn, $N^{\delta}$-hydroxyornithine; β-OH-Asp, β-hydroxyaspartic acid; KB, King's medium B; PTH, 3-phenyl-2-thiohydantoin; TLC, thin-layer chromatography; CD, circular dichroism; *I*, intensity; *F*, structure factor; rms, root mean square; W, water molecule.

0006-2960/81/0420-6446$01.25/0 