The complete nucleotide sequence is reported of the two adult chicken α-globin genes, αA and αD. These two genes, expressed in a 3:1 ratio, respectively, in adult red cells, are widely divergent, suggesting that they have evolved separately for several hundred million years. Although the genes are closely linked in the chicken chromosome, the nucleotide sequences determined clearly rule out any recent gene conversion events. As expected, both genes contain two relatively short intervening sequences. The 3′ intron of the αA gene begins with the dinucleotide GC rather than the typical GT. Extensive flanking sequences are reported for both genes. The chromosomal sequences of the two genes are compared to each other and to sequenced mammalian α-globin genes.

The red blood cells of adult chickens contain two hemoglobins, HbA and HbD, in a 3:1 ratio (1, 2). These two hemoglobins contain an identical β-globin polypeptide chain, but differ in their α components which are referred to as αA- and αD-globin, respectively. The amino acid sequence of the αD-globin protein has been determined (3). We (4) and others (5) have isolated cDNA clones containing nucleotide sequences corresponding to the αD-globin. The nucleotide sequence analysis of these clones is in general agreement with the protein sequence, although some differences do exist (Ref. 4 and see below).

An initial report of the αA-globin amino acid sequence has also been published (6). Several laboratories, however, have failed to obtain a cDNA clone containing the corresponding nucleotide sequence (4–7–10) and have instead obtained α-type cDNA clones whose sequence corresponds to neither of the previously published amino acid sequences for αA or αD. Richards and Wells (8) referred to this sequence as that of a hypothetical “αE-globin,” since they presumed that it must be expressed only upon subjecting chickens to phenylhydrazine-induced anemic stress in order to obtain reticulocytes for globin mRNA purification. We have shown (4), however, that even in cDNA clones prepared from normal chicken red cells the primary α-type sequence is identical with that described by Richards and Wells (8). The protein sequence correspond-

* This work was supported by Grants GM 28837 and HL 24415 from the National Institutes of Health. This is Journal Article 10580 from the Michigan Agricultural Experiment Station. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† Recipient of an American Cancer Society Faculty Research Award.

‡ The abbreviations used are: Hba, hemoglobin A; kbp, kilobase pairs; bp, base pairs; HSV, herpes simplex virus.

EXPERIMENTAL PROCEDURES

Isolation, Mapping, and DNA Sequence Analysis of DNA Fragments—DNA fragments containing the α-globin genes were isolated from λ bacteriophage recombinants (4, 12) as described previously (17). These were inserted into pBR322 plasmid DNA digested with the homologous or identical restriction endonucleases as described (4, 18, 19). Fine structure restriction maps were derived for these subclones using single and multiple restriction enzyme digestion followed by agarose or acrylamide gel electrophoresis (19). DNA fragments to be sequenced were radioactively labeled at 5′ termini using polynucleotide kinase and [γ-32P]ATP or at 3′ termini using avian myeloblastosis virus reverse transcriptase and α-32P-dNTPs as described (20, 21). DNA fragments were then recut with another restriction enzyme and singly end-labeled fragments were purified (17). These fragments were sequenced using the chemical degradation technique of Maxam and Gilbert (20), in some instances using the modifications of Smith and Calvo (21). Sequencing gels (usually 8% acrylamide, 0.4 × 170 × 850 mm) were prepared and run as described (20).

Miscellaneous—Restriction enzymes were obtained from New England Biolabs, Bethesda Research Laboratories, and Biotec, Inc. (Madison, WI) and used according to the recommendations of the manufacturers. Polynucleotide kinase was from Bethesda Research Laboratories or New England Nuclear. Avian myeloblastosis virus reverse transcriptase was provided by J. W. Board of Life Sciences.
RESULTS AND DISCUSSION

Restriction Mapping and Sequencing Strategy of Chicken a-Globin Subclones—The chicken a-globin locus contains three a-type genes whose organization is depicted in Fig. 1A (4, 12). The two adult globin genes are located at the 3' end of locus (relative to transcription) in the order 5'-aD,aA-3'. These two genes were excised from a CharonAA recombinant DNA and subcloned in the plasmid pBRa722 after cleavage with appropriate restriction enzymes (Ref. 4; see “Experimental Procedures”). The primary subclones used for sequence analysis are pBRa7-1.7 containing a 1.7-kbp EcoRI-BamHI fragment including the aD gene, pBRa5-0.8 containing a 0.8-kbp BamHI fragment including the 5' two-thirds of the aA gene, and pBRa6-4.3 containing the aD gene on a 4.3-kbp EcoRI-HindIII fragment. The restriction maps of the subclones were determined by standard techniques (see “Experimental Procedures”) and are given in Fig. 1, B and C, as are the sequencing strategies for the two genes. The sequences of both DNA strands were determined for most of the two genes. Partial sequence analyses which provided definitive identification of the adult chicken a-globin genes have been published previously (4).

DNA Sequence of the aD-Globin Gene—The complete sequence of the chicken aD-globin gene is given in Fig. 2. The coding regions of this sequence are identical with that derived from the cDNA clone described by Richards and Wells (8). Several groups, including our own, have also determined cDNA clone sequences very similar to, if not identical with, the genomic coding sequence (4, 7-10). The few differences among these sequences probably result from genetic polymorphisms in the chickens under study or (more likely) from sequencing errors. The amino acid sequence predicted from the coding sequence (Fig. 2) has recently been verified by Knochel et al. (10). As described under the “Introduction,” this sequence is very different from the original report of the aD protein sequence (6). The source of these differences is not yet known.

The aD gene contains two intervening sequences as previously suggested by electron microscopic R-loop experiments (12). As predicted from the studies of mouse globin genes by Leder and colleagues (22-24), these introns split the coding regions of the aD gene within codon 31 (length of intron 1 = 131 bp) and between codons 99 and 100 (length of intron 2 = 109 bp). The 5' intron probably occurs between nucleotides 2 and 3 of the Arg codon at position 31. While the exact location of this particular splice site is ambiguous from the sequence data, the intron has been positioned such that it is bounded by a GT dinucleotide at its 5' end and an AG dinucleotide at its 3' end in accordance with the structure of essentially all other introns of eucaryotic genes which code for proteins (25).

Consensus sequences which flank the aD-globin gene will be described below.

DNA Sequence of the aA-Globin Gene—The complete sequence of the aA-globin gene is shown in Fig. 3. The coding regions of the aA gene can be identified by comparison to the cDNA clone described previously (4). In the region coding for amino acids 107 to 111, the genomic sequence of Fig. 3 differs considerably from the reported cDNA sequence. In addition, there are seven silent differences in the coding region of the genomic sequence from the reported cDNA sequence which may result from genetic polymorphisms, cloning artifacts or sequencing errors. We have carefully re-examined the cDNA and genomic clone sequence autoradiograms and find only three clear inconsistencies in the cDNA sequence with the genomic sequence in Fig. 3 (CCT at codon 37, TGT at codon 66, and GTG at codon 110). The genomic sequence is probably more reliable since the cDNA sequence was primarily determined on only one DNA strand, and the genomic sequence agrees with the published protein sequences (3, 26) at codon 110. The protein sequence predicted by the aA sequence is shown in Fig. 3. It differs in 2 of the 141 amino acids with the reported protein sequence of aA-globin (isolated...
Fig. 2. DNA sequence of the chicken α^*-globin gene. The sequence of the α^*-globin gene is given. ATA and AATAAA signal sequences are underlined. Numbers below the sequence indicate nucleotide numbering with the predicted (see text) 5' end of the corresponding mRNA as nucleotide 1 (designated cap). The amino acid sequence predicted by the nucleotide sequence is given above the coding regions of the gene along with its corresponding numbering. Y = C or T (when these could not be distinguished on the sequencing gel).
**Adult Chicken $\alpha$-Globin Genes**

CAGGGTGACGCTTGATGGCCCTTCCCAAGCTGGGCGGCACTTGGCAGCTGGCAGCTCGGCTGTGAGGGTCGCAACTCTCAACT

-300 -280 -260 -240 -220

GCGACGGCAGGCGTGGTAGGGACGGCCAGGCCCAGCTGGGCCCAGCTGACGGGCTGGGCCCACGGCCTGCC

-200 -180 -160 -140 -120

CCTCGTGTCAGCGGACACAGCAGGCCCCCTCTGGGATGCCAGGCTGAGGGTCGCAACTCTCAACT

-100 -80 -60 -40 -20

cap

Met Leu Thr Ala Glu Asp Lys Leu Ile Gln Glu Ala Trp

ATCACACATTGCCACCAGCCACCAGCCCGCCCCACCAGCTGCCACC

ATG CTG ACT GCC GAG GAC AAG

AAG

CTC ATC CAG CAG GCC TGG

Glu Lys Ala Ala Ser His Gln Glu Phe Gly Ala Glu Ala Leu Thr Ar

GAC AAG GCC CCT GAC CAG GAG TTT GGA GCT GAG

GCT CTG ACT AG GTGCGAGCCAGCCCAGGGACTGGGCTGGG

Glu Val Arg Gly His His Gln Val Leu Gly Ala Gly Asn Ala Val Lys Asn Leu Ser Gln

CAG GTC GCT GGC CAT GGC AAG AAG ACC TAC TTT CCC CAG ACC AAG

440 460 480 500

TCTCGGCTCTGGGGTCCTGAGCTGCTGGCAGGCTGGGCTGGGCTGG

520 540 560 580 600

GGCTGTGCTGGGGTCCTGGGATCTGGCAGGCTGGGCTGGGCTGGG

620 640 660 680 700

Leu Leu Ser Gln Cys Ile Gln Val Leu Ala Val Lys His Met

CCTCGGCTCAGGACAAAGTGACTGAGCCTCTGTTTGCCTTGCAGG

720 740 760 780

Gly Lys Asp Tyr Thr Pro Glu Val His Ala Ala Phe Asp Lys Phe Leu Ser Ala Val Ser Ala Val Leu Ala Glu

GCC AAA GAC TAC ACC CCT GAA GCT GAT GCC TGC GAG TAC GCT TGC TCT GCC TCT GCT GCT GCT GCT GCT GAG

800 820 840 860

Lys Tyr Arg Stop

AAC TAC ACA TAA GCAACGCTCTACAAGTTGCTGAGACTTCAAGATGCTGAGCTGGGCTGGGACAG

880 900 920 940 960

GGCATGGCGTCAGGGTGCTGGGCACTCAGCTTGGGGCCAGCCCAGCTGATCAGGCTGGGCTGGG

980 1000 1020 1040 1060

TGCCTGGTTAACGCAGGCTGTTTGCAGGACTGAGCCTGGGCTGGGCTGGG

1080 1100 1120 1140 1160

**FIG. 3.** DNA sequence of the chicken $\alpha$-globin gene. The sequence of the $\alpha$-globin gene is given. ATA and AATAAA signal sequences are underlined. Numbers below the sequence indicate nucleotide numbering with the predicted (see text) 5' end of the mRNA as nucleotide 1 (designated cap). The amino acid sequence predicted by the nucleotide sequence is given along with its corresponding numbering.
from embryos) of Chapman et al. (26). Of the six differences between the Chapman et al. protein sequence and the original sequence of Takei et al. (3), the chromosomal nucleotide sequence agrees with Chapman et al. at codons 22, 33, 53, and 54 and with Takei et al. at codon 111, and it indicates a valine codon at position 107, differing from both of the above protein sequences. Although only one DNA strand was sequenced at this codon, its sequence can be read quite easily and is considerably different from the codons for the amino acids (Cys or Gln) previously determined for position 107. Assuming that the amino acid sequences (3, 26) are correct, it may be that codon 107 represents a site at which there exists a high frequency of genetic polymorphism in different chicken lines.

The α\(^{\text{A}}\)-globin-coding sequences are also split by two intervening sequences at the predicted sites. The 5′ intron of 148 bp probably occurs between the second and third base of codon 31 (Arg) just as it does in the α\(^{\text{B}}\)-globin gene (see above); the 3′ intron also occurs between codons 99 and 100. In the α\(^{\text{A}}\)-globin gene, however, the 3′ intron is somewhat larger (261 bp) than is typical for adult α-globin genes. More striking is the fact that the 3′ intron does not conform to the 5′-GT...AG-3′ consensus sequence which bounds essentially all other known intervening sequences of functional eucaryotic protein-coding genes (25, 27), but it rather is bounded by 5′-GC...AG-3′. This sequence has been verified by sequencing both DNA strands in this region (several times), an example of which is shown in Fig. 4. In this experiment the BamHI site located 65 bp into the 3′ intron has been labeled separately at the 5′ and 3′ ends of the DNA. The four sequence ladders of each of the two strands were run side by side, and the region corresponding to the intron donor site (its 5′ border) is shown in Fig. 4. As expected, the two strands are exactly complementary and, in particular, the 5′-GGCAAG-3′ consensus donor sequence with a 5′-GC-3′ at the GT site did not direct splicing in vivo; however, this splicing mutant also completely disrupted the consensus donor sequence 3′ to the GC dinucleotide, so the exact cause of the mutant phenotype remains uncertain. In this regard, as is shown in Fig. 5, the 3′ intron

![Fig. 4. Nucleotide sequence of the 3′ intron donor site of the α\(^{\text{A}}\)-globin gene on both strands.](image)

![Fig. 5. DNA sequence of intron donor and acceptor sites.](image)
donor site of the αβ-globin gene provides an excellent fit to the consensus intron donor sequence (27) with the exception of the C in the GC dinucleotide.

**Flanking Sequences of the α- and αβ-Globin Genes**—Several sequence blocks of DNA that flank eucaryotic genes are presumed to have functional importance because of their widespread occurrence in specific regions of such genes. Perhaps the most distinctive of these is the “Goldberg-Hogness” or “ATA” box (31, 32) which has been found to precede the initiation site of transcription (or, at least, the site of the cap addition to mature mRNA) by about 30 bp. Such ATA boxes can be clearly identified in the chicken α- and αβ-globin genes about 70 bp upstream from the initiator ATG codon (Figs. 2, 3, and 6). These allow us to tentatively place the cap sites (approximate 5' ends of the mature mRNAs) as shown in Figs. 2 and 3. The sequences of the proposed cap sites agree with those determined for other globin genes (Fig. 6A;Refs. 15, 24, and 32) as do their positions relative to the ATA boxes. Further experiments will be necessary to confirm the proposed cap sites for the chicken α-globin genes.

A variety of experiments indicate the presence of at least one other local signal sequence which modulates levels of transcription of eucaryotic genes copied by RNA polymerase I (33, 34). This sequence may include the “CCAAT” regions identified by their presence upstream of the ATA boxes of many isolated genes (32–34). Regions corresponding accurately to the consensus CCAAT cannot be identified in either of the adult chicken α-globin genes. Both genes, however, possess two regions containing limited homology to the CCAAT site. They both contain duplicated pyrimidine-pyrimidine-AAC blocks far upstream from their ATA sites (Fig. 6A). The distance between these sites and the ATA boxes in both cases generally exceeds that observed in other systems (32–34). Regions can also be found in both chicken genes which contain sequences with weak homology to the consensus CCAAT sequence and whose spacing (from ATA) more nearly coincides with typical CCAAT boxes (Fig. 6A). Recently, McKnight and Kingsbury (33) have found that the probable CCAAT sequence is not effective in regulating transcription from the HSV thymidine kinase gene when injected into *Xenopus laevis* oocytes (although the actual location of the CCAAT sequence in this gene is slightly ambiguous due to the incomplete homology of the proposed CGAAT site to CCAAT). Instead, these workers identified two other regions necessary for efficient HSV thymidine kinase gene transcription as a G-rich site (47 to 61 bp upstream from the cap site) and another site (80 to 105 bp upstream) which in part was composed of a C-rich sequence containing inverted repeat symmetry to the above G-rich sequence. The significance of these regions for transcription remains obscure (33), but it is interesting to note that both chicken α-globin genes contain such G-rich and C-rich sequence blocks at appropriate positions upstream from their presumed cap sites (as, indeed, do many other globin genes; Refs. 15, 16, 23, 24, and 32). While the α-globin genes contain several possible examples of these sequence blocks, the most probable G-rich and C-rich sites homologous to those in the HSV thymidine kinase gene are shown in Fig. 6B. It may be noted that it is the case for both genes that considerable inverted repeat symmetry is maintained between the two sites, although this is not totally unexpected when comparing any G-rich block to a linked C-rich block. Further studies will be required to delineate the importance, if any, of these regions to the control of globin gene transcription.

The 3'-flanking regions of the adult chicken α-globin genes also contain sequences common to those observed in other systems.

![Fig. 6. DNA sequence of putative signal sequences flanking the chicken α-globin genes.](http://www.jbc.org)
Adult Chicken α-Globin Genes

Acknowledgments—We are grateful for the excellent technical assistance of Pat Creatura and Sara Stadt in this research.

REFERENCES


Downloaded from http://www.jbc.org/ by guest on September 23, 2017