The complete sequence of the chicken α -cardiac actin gene: a highly conserved vertebrate gene

Kun Sang Chang, Katrina N. Rothblum and Robert J. Schwartz*

Department of Cell Biology, Baylor College of Medicine, Houston, TX 77030, USA

Received 23 November 1984; Revised and Accepted 29 January 1985

ABSTRACT

We sequenced the entire chicken α -cardiac actin gene. A single intron was positioned 20 bp upstream from the initiation ATG codon in the 5' non-coding region while the coding region was interrupted by 5 introns at amino acid positions 41/42, 150, 204, 267, and 327/328. Sequencing allowed the first comparison of the α -cardiac and α -skeletal actin transcriptional promoters. These highly G+C rich promoters share two regions of homology which are found at position -134 (10 bp) and -296 (12 bp) in the α -cardiac actin promoter. A smaller 9 bp motif (CCGCGCCGG) homologous to the -134 sequence was detected before, between and after the TATA and CAAT boxes of the α -cardiac actin gene. The polyadenylation signal (AATAAA) was located 156 bp downstream from the translation termination codon. The complete length of the α -cardiac actin mRNA excluding the poly A tail is 1370 nucleotides. The 3' noncoding transcribed portion of the chicken α -cardiac actin gene was found to be extraordinarily conserved when compared to the human and rat α -cardiac actin mRNA sequences.

INTRODUCTION

The actins represent a multigene family of highly conserved contractile proteins found in a broad host of eukaryotic genomes (1-5). Differences in amino acid sequence between the various actins have shown that at least six different isoforms are expressed in birds and mammals (6,7). The actins found in plants (8), protozoa (9), and invertebrates (3,4,10) are diverse, but tend to resemble the non-muscle β - and γ cytoplasmic actins identified in higher vertebrates (6). The first appearance of a striated actin coincides with the evolutionary emergence of the chordates (11). Urochordates and lower vertebrates such as the lamprey express the α -cardiac actin isoform in their primitive muscles. Vandekerckhove (11) has suggested that the other striated actin, the α -skeletal isoform, appeared during the evolution of amphibia. The smooth muscle actins which are expressed in vascular (α -smooth muscle) and visceral (γ -smooth muscle) tissues were believed to have evolved during the later development of warm-blooded vertebrates. Hypothetically, a primitive cardiac-like actin gene might have given rise to the other striated actin

genes by a series of gene duplication steps. Since none of the striated actins seem to be closely linked to each other at the DNA (1,4,5) or even at the chromosomal (12) levels, several other events must have also transpired which dispersed these genes throughout vertebrate genomes.

We have isolated (1) and now sequenced the entire chicken α -cardiac actin gene. The coding region of this gene was interrupted by five introns which were located in the same positions as introns previously identified in other vertebrate cardiac (13) and α -skeletal actin genes (14,15). Additionally, we have identified a sixth intron in the 5' transcribed noncoding leader of the chicken α -cardiac actin gene. Sequencing of the α -cardiac actin gene has allowed the first comparison of promoter regions between two different striated actin genes. Several short regions of sequence homology were detected in the 5' promoter regions of the chicken α -skeletal and α -cardiac actin genes. Extensive regions of homology were detected in the 3' transcribed noncoding portion and the 3' flanking sequences of the chicken, rat and human α -cardiac actin genes. These highly conserved sequences may provide clues towards understanding the differential regulation of striated actin genes during embryonic development (16,17).

MATERIAL AND METHODS

Materials. Restriction endonucleases and placental RNasin were obtained from New England Biolab and Promega BioTec. AMV-reverse transcriptase was purchased from Life Sciences Inc., and *E.coli* DNA polymerase I large (Klenow) fragment was bought from Bethesda Research Laboratories. Radioactive nucleotides were obtained from Amersham Inc. T4 DNA ligase was a product of Collaborative Research Inc., while nucleotides and the 17bp M13 sequencing primer were from PL Biochemical.

DNA Restriction Mapping A 7.5 Kb *Eco* RI fragment which encompassed the entire cardiac actin gene was isolated from λ AC7, a chicken genomic λ Charon 4A clone (1), and subcloned into the *Eco* RI site of pBR322. The restriction endonuclease enzyme map of pAC 7.5 was determined by single and double enzyme digests. Maps were also developed by the partial digestion protocol of Smith and Birnstiel (18). Reaction conditions for restriction endonuclease cleavage of DNA varied with the particular enzyme and were essentially as suggested by the supplier. Digestion products were separated by agarose slab gel electrophoresis in buffer containing 50 mM Tris-HCl (pH 7.4), 20 mM NaOAc, 18 mM NaCl, and 2 mM EDTA at 40 volts, blotted into nitrocellulose filters and hybridized with [32 P]-labeled DNA probes as previously described (1).

Nucleic Acid Sequencing. The 5.6 Kb Kpn I/Eco RI fragment of pAC 7.5 which contains the entire actin coding region was mapped with a variety of restriction enzymes as shown in Figure 1. Individual DNA fragments were purified by preparative agarose gel electrophoresis. Staggered ends of the DNA fragments were filled in using the Klenow fragment of DNA polymerase I and deoxynucleoside triphosphates. The blunt ended fragments were ligated into Sma I cut M13mp10 in both orientations. Two small overlapping M13mp10 libraries were constructed from a 1.9 Kb Kpn I/Pst I fragment which was digested with Sin I and Dde I. The 0.7 Kb Sma I/Kpn I fragment was ligated in opposite orientations into mp18 and mp19. Identical clones were identified by T sequencing tracks, while clones containing DNA inserts of the opposite orientation were identified by the C-test (19). Single stranded M13 clones were sequenced by the dideoxy chain termination method of Sanger (20,21). The sequencing reaction mixtures were electrophoresed on preheated 6% polyacrylamide gels containing 8M urea (21). Gels were autoradiographed with Kodak AR film at -20°C for 4-12 hours. DNA sequences were compared to the published α -cardiac actin amino acid sequence (22) with the aid of an Apple II plus computer and the software of R. Larson and J. Messing (University of Minnesota).

Nucleotide Sequence of the 5' end of α -Cardiac Actin mRNA. The Sin I fragment cloned into M13 which overlaps the translational start site was isolated from RF M13 DNA and digested with Xma III. A 140 bp fragment was isolated, blunt ended with Klenow and subcloned into the Sma I site of M13mp10 to produce clone XS5. The 17bp sequencing primer was end labeled with [γ^{32} P]-ATP and T4 polynucleotide kinase, and then used to extend a DNA chain in the presence of the single stranded clone XS5, 0.25 mM dNTP, 5 mM DTT, 10 mM MgCl_2 , and 50 mM Tris-HCl (pH 7.6). The labeled extension product was digested with Hha I to produce an 84 nucleotide fragment which was complementary to a portion of the 5' end of the α -cardiac mRNA. This end labeled product was denatured in 30% DMSO and isolated on a 5% polyacrylamide gel. The single stranded labeled primer was hybridized with 100 ug of chicken cardiac poly A⁺ RNA, isolated as described in Chang *et al.* (1), in 40 ul of 25 mM Pipes (pH 7.5), 0.4 M NaCl and 50% formamide at 50°C for 24 hours. Hybridization was terminated by ethanol precipitation and the pellet was washed three times with 70% ethanol, once with 100% ethanol and then lyophilized. The dry pellet was resuspended in a primer extension reaction mixture which contained 10 mM DTT, 50 mM Tris-HCl pH 8.3, 0.12 M KCl, 7 mM MgCl_2 , 0.25 mM dNTP, 20 units RNasin, and 25 units of AMV-reverse transcriptase in a 40 ul volume and was incubated for 1

hour at 42°C. The primer extended product was gel purified and sequenced by the Maxam and Gilbert protocol (23).

Location of the 3' end of the α -Cardiac Actin mRNA. In order to make an S_1 nuclease probe to the 3' end of the α -cardiac actin mRNA, the 1.2 Kb Eco RI/Sin I fragment located 30 bp from the translation termination codon was blunt ended and cloned into the Sma I site of M13mp10. A clone was selected in which the destroyed Eco RI site of the insert DNA was proximal to the M13 primer. A [32 P] labeled DNA was synthesized by primer extension, cleaved at the Eco RI site in the M13mp10 cloning cassette, denatured, and recovered as a single stranded probe following preparative electrophoresis. This probe was hybridized with 10 ug of cardiac poly A⁺ RNA as described above. The S_1 nuclease reaction was performed in a final volume of 250 ul which contained 0.2 M NaCl, 30 mM Na acetate (pH 4.5), 5 mM ZnSO₄, and 20 ug of yeast t-RNA at 37°C for 1 hour in the presence of several concentrations of S_1 nuclease. The samples were extracted with equal volumes of phenol-chloroform, and coprecipitated with carrier t-RNA (5 ug) in 70% ethanol. The size of the protected fragment was determined by comparison to thymidine sequencing ladders of an M13 clone which was co-electrophoresed on the same 6% polyacrylamide slab gel.

RESULTS AND DISCUSSION

Our laboratory has previously reported the isolation of the single copy α -cardiac actin gene from a λ Charon 4A chicken genomic library (1). The 7.5 Kb DNA fragment containing the coding and flanking sequences was isolated from λ AC7 and subcloned into the Eco RI site of pBR322 (pAC7.5; Figure 1). The detailed map of pAC 7.5 shown in Figure 1 guided our preparation of DNA fragments for cloning into M13 vectors. From these clones we determined the 11-near sequence of 5463 nucleotides of the chicken α -cardiac actin gene as shown in Figure 2.

The Amino Acid Coding Region of the α -Cardiac Actin Gene.

Sequencing of the α -cardiac actin gene revealed the presence of 377 encoded amino acid residues, which began with a Met-Cys dipeptide (Fig. 2). This dipeptide is also found in the human α -cardiac actin gene (13) as well as the other vertebrate striated actin genes (14,15). Interestingly, this dipeptide is found in all of the Drosophila non-muscle actin genes (10) and some sea urchin actin genes (3) while the Cys codon is absent in vertebrate non-muscle actin genes (24,25). Since the NH₂-terminal Met or Cys has not been detected in an actin protein yet sequenced, it is assumed that post-translational re-

Figure 1. Structural map of the chicken α -cardiac actin gene. Genomic DNA (7.5 Kb) which circumscribes the actin coding region was subcloned into pBR322 (pAC 7.5) and then mapped with the following restriction enzymes: Dde I (D); Hind III (H3); Kpn I (K); Pst I (P); Pvu II (P2); Sau 96I (S9); Sca I (Sc); Sna I (Sn); Sma I (Sm); Eco RI (RI); Sac I (SI); Xba I (Xb); Xho I (X2). These maps led to a sequencing strategy in which fragments were subcloned into M13 vectors and sequenced by dideoxy chain termination as indicated by arrows. One squiggly arrow represents a fragment sequenced by the Maxam and Gilbert (27) protocol. In the diagram above, the solid boxes represent coding exons, open boxes indicate transcribed untranslated regions and the solid lines coincide with introns and flanking DNA. The TATA box, CAAAT box, initiation codon (ATG), termination codon (TAA), and the poly A addition signal (AATAAAA) are indicated.

moval of the Met-Cys dipeptide might be required for normal actin function (15).

The amino acid sequences at the NH_2 - and COOH terminal regions provide definitive identification of each vertebrate actin isoform (7). The Asp and Glu at positions 2 and 3 as well as the Leu and Ser at positions 298 and 357 provided the 4 amino acids which differentiate the α -cardiac actin isoform from that of the α -skeletal type. We determined that the nucleotide sequence encoded a chicken protein which is absolutely identical to the bovine and human (13,22) α -cardiac actins. Thus, the sequence of the α -cardiac actin isoform has been strongly conserved during the evolution of warm-blooded vertebrates.

The coding region of the α -cardiac actin gene was split by five introns at the codons specifying amino acids 41/42 (IV 2), 150 (IV 3), 204 (IV 4), 267 (IV 5) and 327/328 (IV 6). The length and location of the introns are summarized in Table I. The sequences of the intron-exon junctions agree quite well with the consensus sequences compiled from the study of many genes (26). The positions of splice junction boundaries in the coding region are identical to those in the human α -cardiac actin gene (13) and the α -skeletal actin genes (14,15) isolated from chicken and rat. Unfortunately, only partial sequences of the human α -cardiac actin gene are available for

-280
 CCGGGGGCCGACATCTCGGCGACGACCGGCTTGTACATTCCGGAGGGCCATGGAGGGCCACAGAGACTCCGAGCGCTCCC
 -200
 CTCGCCACCCAGCTGGCGGGGAGGCTCCCTATTTGGCCATGTGGCGGCGGGACCGGTGCTCTCACTGCGCTTAGATGGCC
 -120
 GGACAGCGCGCGCGGCTTGGCCATTATGGCCGCGCTGCGCGCATGGCGAAGCGGCAAAATAGGAGAAGGTGGCTGCC
 -40
 CGGGCGGGGACGCGCGGCGCGGGGCTATAAAGCGGCAGCTGCGCGCCCGCGGACCGGCGCTCTCGGTAACC
 TCCCGCGCTCCGTGCTCCCG/GTGAGTGTGTGCTGTAAGCGGGGACGGGAAGGATGGAGAGAGAGAAGATGGATGGG
 GGGCACCGACGGGAGGCGACCGGCTACGTTGTGCGGTGTATCCGCGCTCCGCTGCTCTGCGCGAGACGGGA
 GGGGACAGCGGCTCTGCGCTTCTCGGCTTCGGAGCGGTGCGGCTGACCGGGGCGGGTGGCGCGGAGGCGGCGTGGC
 CTGGCAGAGCGGCACTTCGGCATACCATTTGCTCATCCATTTTGTATCTGCCCCATAGGGTACCGGTAGTGTATGGCCCA
 AAGTACCGGTTCTGGGAGCCTTCTGAATCGTGAACGCGGCGGCGGCGGCTGCTCAGAGCGTGGCCGAGCATGGCCG
 AGCCCCGCGGACCGCGGAGCCACCGCTGCCACCGGGGCTGCCCTGTCCCGGGGTGCCCGAGGTGCCCGCAGAG
 GTCTGTATGCGCGCGCGCTGTCTTGAG/(820)

Met Cys Asp Asp Glu Glu Thr Thr
 TGT TGT GAC GAC GAG GAG ACC ACC
 1 5
 10 15 20 25
 Ala Leu Val Cys Asp Asn Gly Ser Gly Leu Val Lys Ala Gly Phe Ala Gly Asp Asp Ala
 GCG CTG GTG TGC GAC AAC GGC TCG GGG CTG GTC AAG GCG GGC TTT GCT GGG GAC GAT GCC
 30 35 40 45 50 55
 Pro Arg Ala Val Phe Pro Ser Ile Val Gly Arg Pro Arg His Gln
 CCC AGG GCC GTG TTC CCA TCC ATC GTC GGC CGC CCC AGG CAC CAG /GTAAATCTCTGCCCGATG
 ACAGAAACGGGTCCATTCTCTCATCAGGAGTCAGCCCACTCTTAATCTGTCTCTGCCACCCCGAGCTCTCT
 CTTCCTTGTGCTCTACTGTACACATTTCTGTGTCTGTGCTTCTTCTTCACTTCCCTCTTTGGCTCACCATCATCT
 TATCTCTGTCACTGTCAACCCCTGTACCACTTCTTTCCCAATTCACCAAGTATCCCTTCTCTCTCCCAATGAGTCCAA
 ACCAATTTCTCTCAGAACTGCTTCCCAACACCTCCCTAAGTGAGATAAGTCTGCGTGGTGTGGTCTAAAGACATCTCC
 ATGGGCTGAATTTGTCCACCCATCTCCATGCCCATATAAGCCCTTGCAATTTCTCTGCTGCTTTGCTGTCTGTCTG
 CCTTCTTACCACTGTCTAAGCGTTCTTTCTGGGGTCCCCATTTCACTTCTCTGTTATTTCTCTCAAGTTGTCTTCAAA
 CACAGAAACGTAAGGCTTTCTTTTATAATTTCTGGCAAAAACTTATCATGCTCTTTTCACTGTCAATGAACTCT
 GGTCCAGAAAGAGGTAATTGAAATAAAGGATATATTTAAAAAAAATAGTGGGAAAAGTCCCTTTGATCTGATTAAAGGA
 AAAATAACTTTGAGAACTTCAAGTCTGTGTGAGTAAATCTTATAAACCACTTTTAAACACCTTTTAGTGGCCCACTAC
 TGTGTTCACTCTCTGAAATAAGCATGTTGATTGAGGATAAATCTAGCTCTGTGGTGATGCTTTCTTCTCAAACTATG
 CGTTCTCTGATTTTGTCTTTTGA/(1801) 45 50 55
 Gly Val Met Val Gly Met Gly Gln Lys Asp Ser Tyr Val Gly
 GGT GTT ATG GTC GGT ATG GGT CAA AAA GAT TCC TAT GTG GGT
 60 65 70 75
 Asp Glu Ala Glu Ser Lys Arg Gly Ile Leu Thr Leu Lys Tyr Pro Ile Glu His Gly Ile
 GAT GAA GCA CAG AGC AAA AGA GGT ATC CTT ACT TTG AAA TAT CCT ATT GAA CAT GGC ATC
 80 85 90 95
 Ile Thr Asn Trp Asp Asp Met Glu Lys Ile Trp His His Thr Phe Tyr Asn Glu Leu Arg
 ATC ACC AAC TGG GAC GAT ATG GAA AAG ATC TGG CAC CAC ACT TTC TAC AAT GAA CTC CGT
 100 105 110 115
 Val Ala Pro Glu Glu His Pro Thr Leu Leu Thr Glu Ala Pro Leu Asn Pro Lys Ala Asn
 GTG GCC CCT GAG GAG CAC CCC ACC TTG CTC ACA GAA GCA CCC TTG AAC CCT AAA GCT AAT
 120 125 130 135
 Arg Glu Lys Met Thr Gln Ile Met Phe Glu Thr Phe Asn Val Pro Ala Met Tyr Val Ala
 CGT GAA AAG ATG ACC CAG ATA ATG TTT GAG ACC TTC AAT ATA CCC GCT ATG TAT GTG GCA
 140 145 150 155
 Ile Gln Ala Val Leu Ser Leu Tyr Ala Ser Gly Arg Thr Thr
 ATC CAG GCT GTC CTG TCC CTG TAT GCC TCT TGA CGT ACC ACA G /GTAAAAATAGACTTCTTTCTCT
 TTCAGACCCACAGACCCCTCTCTCTCCCACTCCAAACACACAGACACACAGACACAGAAACACGTCGATGTCTCAAAT
 TTACAGAAAACTATGCTGTTGGCTTTTTTATTATTATTATTATTATTAAAGAGTAACCTAGCAGAAAAAGCTAGCTGTG
 TGAAGAACTATAATAATTTGTCTGATTGTTGTTTAAAAAGAAAGGCACTTAAAGAACCTGTATTTTCTGCTGCTGTGATC
 GTGCAAAACAGGAGACATCACTGCACTGAGCAAGTATTACACCACTTTGAGTTTAACTGAGCCTAAGAGTATGGCTT
 TTTACTGTATCAAGCGGTTGCTACCATCAATACACTTGCTGCTCAGTGTAGCATTTGAGGCTATTCAATATGATGATCA
 TTAGAGCTGTGATTTGGGAAATTTTTATGGGATTTTACTACCAATGGGAATTTTGAAGAGTGTGGCAAGCTGCA
 ATGAGCAGTAGCTGGCTAAAAATAGAAATGACGCTGCTTATGCTGCTGAAATGCAATGTCTGATATCTGTGGTATGCTGTG
 GTTGAAGGTTTCTCAGGTTATTCCAATGGTCTTCTGACATTTCTTTCCAG / 150 155
 (2751) Gly Ile Val Leu Asp Ser
 GT ATG TCT CTT GAT TCT
 160 165 170 175
 Gly Asp Gly Val Thr His Asn Val Pro Ile Tyr Glu Gly Tyr Ala Leu Pro His Ala Ile
 GGT GAT GGA GTC ACC CAC AAC GTG CCC ATT TAT GAA GGT TAT GCT TTG CCT CAT GCT ATC
 180 185 190 195
 Met Arg Leu Asp Leu Ala Gly Arg Asp Leu Thr Asp Tyr Leu Met Lys Ile Leu Thr Glu
 ATG CGT CTG GAT CTG GCT GGC AGG ATC CTG ACC GAC TAC CTC ATG AAG ATC CTC ACA GAA
 200 205 210 215

200
 Arg Gly Tyr Ser Phe Val Thr
 CGT GGC TAT TCC TTT GTG ACC ACA G /GTCAGTCCCTTCAGCTTGACCCTCTTCTGAATATCCTCATGTCTC
 CCTTGGCCACCTTCTTCCCTGATATTTTGCTTCTCCAACTTGATTGATTCTGTGTACCTGCAG / 205
 (3025) Ala Glu
 CT GAA
 225
 Arg Glu Ile Val Arg Asp Ile Lys Glu Lys Leu Cys Tyr Val Ala Leu Asp Phe Glu Asn
 CGT GAG ATT GTT CGT GAT ATC AAA GAG AAA TTG TGC TAT GTT GCG CTG GAC TTT GAG AAT
 230 235 240
 Glu Met Ala Thr Ala Ala Ser Ser Ser Ser Leu Glu Lys Ser Tyr Glu Leu Pro Asp Gly
 GAA ATG GCC ACA GCT GCT TCG TCA TCC TCC CTG GAG AAG AGC TAT GAA TTG CCT GAT GGT
 245 250 255 260
 Gln Val Ile Thr Ile Gly Asn Glu Arg Phe Arg Cys Pro Glu Thr Leu Phe Gln Pro Ser
 CAA GTG ATC ACA ATT GGT AAT GAA CGT TTC CGC TGC CCG GAA ACC TTA TTC CAG CCA TCC
 265
 Phe Ile
 TTT ATT G /GTAAGTGCTTCAAGGAAATTCCTTTTTCAGTAGGAGGTGCAAAATGAATCAGAAGTTCTAGAAGGTGTC
 CCACTGCCCTCTCTGTTTATCTACATTTTGCACAAAATCAACTATAAATAAGGAGTTTGATTCTCTGCTCCCTGTTTTC
 TGTGCTTAATGTGATTACTAGCTAATAATAGCTGATCTATGGACTCCTGTTGCAACTCTAGGGCCAGCTTTTAAAGTTC
 ACAATTTTATGATGGAATGTACACTGACAAATGTGCAATGGCAATGGGCAAGAGCAGCTTTCAGTGCCTGATTTGTGA
 CCACCTGCATTTCAAACACTATCTGATTCCATGAAGAAGTAAATGTGTGGGAGATAAAAGATGTAATGAAGAAGTGT
 ATATTTACATGCCTGAAATTAATCTTGCCGTGGACAAGATGGCTGGTTTGTCTGTGTAGATGCAGCTGACAGAAC
 AGCAATGCTTATAATCTAAGCCCAAGCAAAATGCAGTTGTTGCTCAGCTGATGTGTTACTGTGGATTAGTGTCTAG
 TTCTTTCAGGAAGCACCTGTTAAAAATATAAATTTGGAATGCAACCTACGAACACTTGCTTCTGGTTTCTTTTACTTAT
 CACCAATGCTAACATCTGATCTATGCTCTTGGCTTGTCTTCTCATGTTTATAAATGGCAGATAGTCTACAATAAATAGT
 TTATCTTATGTGCTAACTCTTCTTACTGTAGGCTTGTGTCTTTTCTTGCACATGCATTATGATTCTTAGCTTTT
 CTAATAAATAATATTAGTTTCTAAGCTTGTCTATAGTCTGCGTTGTTTTTGGGTTTTTTTTTTTTTTTTTTTGTAAT
 TTCTGTACAACCTTAGTAATCAATAAGCAATTTTGAACAAAACAATAAAAGTTCATATTATAGTGAGGAGAAATGCT
 GAGTGTCTTTTTCACAAAGACCTCATGTATAATTTACTGCATCAGCTTCTTTTCTCAAG / 270
 (4215) Gly Met Glu Ser
 GT ATG GAA TCT
 290
 Ala Gly Ile His Glu Thr Thr Tyr Asn Ser Ile Met Lys Cys Asp Ile Asp Ile Arg Lys
 GCT GGT ATT CAT GAA ACT ACC TAC AAC AGC ATC ATG AAG TGT GAT ATT GAT ATT CGT AAG
 295 300 305 310
 Asp Leu Tyr Ala Asn Asn Val Leu Ser Gly Gly Thr Thr Met Tyr Pro Gly Ile Ala Asp
 GAT CTG TAT GCT AAC AAC GTC CTC TCA GGT GGT ACA ACA ATG TAC CCT GGT ATT GCT GAT
 315 320 325
 Arg Met Gln Lys Glu Ile Thr Ala Leu Ala Pro Ser Thr Met Lys Ile Lys
 CGC ATG CAG AAA GAA ATC ACT GCA CTT GCT CCT AGC ACA ATG AAG ATC AAG /GTGAGGACTT
 TGATTCTACTAAAAATTTGAATAAGGTTTAGGAAAATGTCATTAAATAGGATACAGGAACAACCTCATAAATGAGACC
 ATCAGCTGATGGCGCGTCCGGCCCAAGAGAATAATTTTCATTTCACCTTCATTAAATATTAAGATATTAATTTGAAGCAA
 TGTGCAATCAGGAAAATGATCTGTAGATTTAGTTAAAGAATATTACATTACCAGCTGCTGAAGACCAGTACATTACT
 ATACATTATTACTTTAAGAAACACCTAGCAGTTTGGTGTGTGATGATGTGGGTAGTTGTGCGAGGAATGAACCTGTTA
 ACGTAGGATTTGGCTGTTTCTCTTCCTAG / 330
 (4756) Ile Ile Ala Pro Pro Glu Arg Lys Tyr Ser Val
 ATC ATT GCT CCA CCT GAA CGC AAG TAC TCT GTC
 340 345 350 355
 Trp Ile Gly Gly Ser Ile Leu Ala Ser Leu Ser Thr Phe Gln Gln Met Trp Ile Ser Lys
 TGG ATT GGT GGC TCC ATT CTA GCA TCT TTG TCC ACC TTC CAG CAA ATG TGG ATC AGC AAA
 360 365 370 374
 Gln Glu Tyr Asp Glu Ala Gly Pro Ser Ile Val His Arg Lys Cys Phe ***
 CAG GAG TAT GAT GAG GCT GGT CCA TCC ATT GTC CAC CGC AAG TGC TTC TAAGACTTTTCCCTA
 CATTTACTTCCCACTCAGGATGATGACAGTATGCTTCTTGGAGTCTTCTGGCAGCCCTTCTGAACTCCTCCGTAAGT
 TACAGTTTGTATACACATTTGTCAATTTATTTGTGCTTCTAATATTATGCTTTATAAATAAAGATCTGGACTTGCAA
 CCTACAAATGTTTCTGATTCTTTTATTCTAAGCGTTATTGGTTAAAAATGAGTCTGTGAATACCTGGATTGGAAACT
 CCCATCAGATTTGTCTACGTGCTCTTCTAAGTGATGTGGACATGTTAGGTCATATTTTCATATTTATAGTCTTAATTTG
 TAAACTTGAAACAAAAGAACTATCATTTACAAGAAAGTAACCTTGAGGGAAGACCAAGGAGTGAACAAGAAGGAAA
 CCTCTCACATTCACTGAATGGGCTTAGAATTGCTTTTTTTCATTGCGTTAATTATAACCTCAGTAATTTAATATCAAC
 TTACTGAAAGCCAAAGCCTCTCAAGAAGGATTATGCTAATACCTAGAAAGCAATTTATACTAATTAACCTCAGT
 (5463)

Figure 2. Nucleotide sequence of the chicken α -cardiac actin gene. Negative numbers indicate nucleotide positions upstream from the transcription start site. Amino acids of the actin protein are indicated above their respective codons, and numbered according to Lu and Elzinga (35). A slash represents a splice junction border, while numbers in parentheses indicate total number of nucleotides sequenced. The CAAAT box, TATAAA box, and the final 3' nucleotide in the mRNA are underscored. The cap site is indicated by a number 1.

Table I
Organization of the Chicken α -Cardiac Actin Gene

Exon/Intron	Positions (number of nucleotides)	Size (nucleotides)	Encoded region of mRNA
Exon I	1-40	40	40bp 5' untranslated region
Intron 1	40-522	482	-----
Exon II	522-671	149	20bp 5' untranslated region to AA 41
Intron 2	671-1503	832	-----
Exon III	1503-1828	325	AA 42 - AA 150
Intron 3	1828-2453	625	-----
Exon IV	2453-2615	162	AA 150 - AA 204
Intron 4	2615-2727	112	-----
Exon V	2727-2919	192	AA 204 - AA 267
Intron 5	2919-3917	998	-----
Exon VI	3917-4099	182	AA 267 - AA 327
Intron 6	4099-4458	359	-----
Exon VII	4458-4778	320	AA 328 to 3' untranslated region

comparison (13), and no extensive sequence homology was detected between human and chicken α -cardiac actin introns.

The 5' Untranslated Region. The sequence of the 5' untranslated region of the α -cardiac actin mRNA was determined by a primer extension experiment summarized by the schematic diagram in Figure 3. The [^{32}P] end labeled

B.

Figure 3. Sequence determination of the 5' end of the chicken α -cardiac mRNA. The autoradiograph of the Maxam and Gilbert sequencing gel in panel A shows the 5' leader sequence of the cardiac actin mRNA. Nucleotides within the coding portion of the mRNA are divided into codons. The presumptive splice junction (sp), end labeled primer (P), and the starting point of the primer extension reaction (PE) are shown. The schematic diagram in panel B summarizes the procedure for synthesizing the primer extension probe. The 140bp Xma III/Sin I fragment was isolated, blunt ended with Klenow and ligated into M13mp10 vector at Sma I site. Single stranded DNA from this clone was annealed with end labeled M13 17bp primer and extended with Klenow. The double stranded DNA was restriction digested with Hha I, the Hha I/Xma III/17bp M13 primer fragment was separated in a strand separating 5% polyacrylamide gel; the probe was then hybridized with α -cardiac actin mRNA. See Material and Methods for a detailed description of this procedure.

fragment (84 nucleotides) which spans the 9th (Leu) to the 36th (Ser) amino acid of the actin coding region was hybridized to chicken heart poly A⁺ RNA, extended with reverse transcriptase, and then sequenced. The first lane of the sequencing gel shown in Figure 3 contains the non-extended primer while the second lane contains major and minor extension products of 87 and 85 additional nucleotides which may represent respectively, primary and secondary transcription start sites. The remaining lanes contain the extended primer cleaved with sequencing reagents. The sequence derived from this gel is shown beside it. The primer extension sequence exactly overlaps the coding sequence of α -cardiac actin by 27 nucleotides. Comparison of the extended sequence to that of the gene showed that an intron interrupted the 5' untranslated region 20 nucleotides upstream from the initiation methionine codon. This 5' intron must continue past the Kpn I border, since the remaining 39 nucleotides of the extended 5' untranslated region were not detected within the subcloned 5.6 Kb Kpn I/Eco RI fragment.

The location of the 5' non-coding exon of the α -cardiac actin gene was determined by the hybridization of labeled extension products (synthesized as described in the legend to Fig. 3 and Methods) to a blot of restriction enzyme digested pAC 7.5 as shown in Figure 4 panel A. In panel B the same fragments were hybridized with the non-extended DNA probe. The 0.8 Kb Sma I fragment, and the 0.7 Kb Sma I/Kpn I fragment were found to be differentially labeled by the extension products (lanes 1-3). These results directed us to sequence the Kpn I/Sma I fragment. We assigned the 5' and 3' borders of the first intron by matching the sequence of the 5' untranslated region shown in Figure 3 with that of this fragment. The signals associated with accurate initiation of transcription were identified 547 nucleotides upstream from the initiation ATG. A canonical promoter sequence (TATAAA) was found an appropriate 27 nucleotides upstream from the major extended sequence. A CAAAT box was located 79 nucleotides 5' to the transcription start site. The positions of these regulatory sequences correspond well with the positions of similar sequences found upstream from the 5' cap site of other eukaryotic genes (27).

Since the α -cardiac and α -skeletal actin genes appear to be closely related and co-expressed during embryonic cardiac and muscle development (16,17) we investigated the possibility that they might also share some common nucleic acid sequences in their flanking regions. Upon comparison, several interesting features of these regions became apparent. First, the chicken α -cardiac and the chicken and rat α -skeletal actin genes contain extraordinarily G+C-rich domains which encompass the transcriptional promoter regions. Within

Figure 4. Location of the 5' noncoding exon of the α -cardiac actin gene. One μ g of plasmid DNA was digested with 5 units of *Sma* I/*Eco* RI (1); *Sma* I/*Kpn* I (2); *Sma* I (3); *Hha* I (4); *Pst* I/*Eco* RI (5). DNA fragments were electrophoresed in agarose and transferred bidirectionally onto 2 sheets of nitrocellulose. Filter (A) hybridized to the end labeled primer extended cDNA of cardiac mRNA, and filter (B) was hybridized to the end labeled primer as indicated in Figure 3B.

the chicken α -cardiac gene a region extending a total of 298 nucleotides upstream of the cap site and another region of 522 nucleotides downstream of the cap site had base compositions of about 70% G+C, while the 100 nucleotides immediately preceeding the transcriptional start site was 74% G+C. This G+C richness contribute a striking feature to the striated actin genes: an abundance of potentially methylatable sites (27).

The comparison of the nucleic acid sequences in the promoter regions of the α -cardiac and α -skeletal genes revealed several homologous sequences. The two longest homologous sequences were found at positions -296 (CCGGCCCCGACA) and -134 (GCCGCGCCGG) in the chicken α -cardiac actin gene. In the chicken α -skeletal actin gene these two sequences comprise a single sequence (GCCGCGCCGGCCCCGACA) found between positions -92 and -109. In the chicken α -

Figure 5. Endonuclease S₁ mapping of the polyadenylation site. The autoradiograph in panel A shows the S₁ nuclease protected fragment. Uniformly labeled 3' specific probes were hybridized to α -cardiac mRNA. Hybrids were treated with S₁-nuclease at various concentration from 500 to 3000 units. The [³²P]-labeled 3' specific probe without mRNA and carrier t-RNA (lane c) was treated with 1000 units of S₁ nuclease;; A thymidine-sequencing ladder of an M13 clone served as size marker in lane M. In panel B is a brief restriction map of 3' region of the α -cardiac actin gene illustrates the size of the probe (P) used in this experiment and the reduced size of the S₁ nuclease protected fragment (PF).

cardiac gene the homologous sequence located at position -134 was reiterated at positions -17, -35, and -110 as a series of nine nucleotide direct and indirect repeats (consensus sequence of CCGCGCCGG). It is conceivable that these short nine base pair repeat sequences could potentiate transcription in the α -cardiac actin gene by either forming a stem/loop structure (positions -4 to -52) or by acting as separate regulatory elements. This hypothesis is currently being tested in our laboratory by *in vitro* transcription assays utilizing the *Xenopus* oocyte system.

The 3' Untranslated Region. A potential polyadenylation signal (AATAAA) was identified at nucleotide 5050 of the sequenced gene (Fig. 2). S₁ nuclease mapping (Fig. 5) indicated that the polyadenylation site was located 20

```

TGAGGCGG-----CTACCTTACA-CTCAGGACGACAGTATTGTGCTT---GGAATCCTCTGAGG   R
TAAGATGCCTTCTCTCTCCATCTACCTTCCA-GTCAGGATGACGGTATTATGCTTCTTGGAGTCTTCCAAAC   H
TAAGCA--CTTTTCCCTACATTTAC-TTCCCACTCAGGATGATGACAGTATGCTTCTTGGAGTCTTCTGGCA   Ck

```

```

-C-CCTCTCC-CATCTCCCATCAGTCATTGTACAGTTTGTTTACACACGTGCAGTTTGTTTGTGCTTCAAAT   R
-CACCTTCCCTCATCTTTTCATCAATCATTGTACAGTTTGTTTACACACGTGCAATTTGTTTGTGCTTCTAAT   H
-GCCCTTCCT-GAACTCCTCCGTAC-TTGACAGTTTGTTTACACACGTGCAATTTATTGTGCTTCTAAT   Ck

```

```

ATTATTGCTTTTATAAATAAACAGGCCAGG-CTTGCAACCT-----R
ATTATTGCTTT-ATAAATAAACAGACCAGGACTTGCAACCTATAAAAGCCTCTCGTTTGTGTTTGGGGTA   H
ATTAT-GCTTT-ATAAATAAAGATCT---GGACTTGCAACCTAATAAATGTTTCT-GATTCTTTATTCTAA   Ck

```

Figure 6. Comparison of the 3' non-coding region of α -cardiac actin genes. The sequences have been aligned (indicated by dashes) so that maximal homology was obtained between rat (R), human (H) and chicken (Ck) cardiac actin genes. Regions of extensive homology between the three species were blocked in.

nucleotides downstream from this site. Therefore, the 3' untranslated region of α -cardiac actin mRNA is 179 nucleotides long and the complete length of the messenger, excluding the poly A tail, is 1370 nucleotides. In comparison, the 3' untranslated region of α -skeletal actin mRNA is about 75 nucleotides longer than the α -cardiac actin mRNA (14). There is only limited sequence homology between the 3' untranslated regions of the chicken α -cardiac and the other actin genes (29,30). However, one region of 12 nucleotides (CACTCAGGATGA) located 30 nucleotides past the termination codon has been conserved in the 3' untranslated regions of α -cardiac and α -skeletal actin mRNAs in chickens, rats and humans.

The most striking feature of the 3' untranslated region was the high degree of sequence conservation which is shared among α -cardiac actin genes in warm-blooded vertebrates. This observation was first made by Mayer *et al.*, (29) for rat and human species and we have extended it to the chicken. Two major parts of the 3' untranslated region which are 48 and 92 nucleotides in length showed 84% and 98% sequence homology, respectively (Fig. 6). The overall homology of the 3' untranslated region of the α -cardiac actin genes in these three vertebrate species is about 80%, which is comparable to the overall sequence homology (84%) found in their coding regions. Comparison of the 3' untranslated regions in various vertebrate α -skeletal and β -actin mRNAs has also shown some regions of sequence homology within each of these actin isoforms (29,30). In comparison, other vertebrate genes such as the rat and

human insulin genes (31,32), or the prolactin genes (33,34) which also share common coding regions usually contain greatly divergent 3' untranslated regions. Presently, the role of highly conserved sequences in the 3' untranslated region of the cardiac actin gene is unknown, but the preservation of these sequences for over 200 million years of evolution suggests an important biological role for the expression of the α -cardiac actin genes. Currently, gene transfer experiments are being performed in our laboratory to investigate the biological function of these conserved sequences in the 5' promoter and the 3' noncoding regions of the chicken α -cardiac actin gene.

ACKNOWLEDGEMENTS

This work was supported by Public Health Service grants NS-15050 and a grant from the Muscular Dystrophy Association. R.J.S. is a recipient of a USPHS Career Development Award.

*To whom correspondence should be addressed

REFERENCES

1. Chang, K.S., Zimmer, W.E. Jr., Bergsma, D.J., Dodgson, J.B., and Schwartz, R.J. (1984) *Molec. and Cell Biol.* 4, 2498-2508.
2. Fyrberg, E.A., Kindle, K.L., Davidson, N., and Sodja, A. (1980) *Cell* 19, 365-378.
3. Scheller, R.H., McAllister, L.B., Crain, W.R., Durica, D.S., Posakong, J. Thomas, T.L., Britten, R.J. and Davidson, E.H. (1981) *Mol. Cell. Biol.* 1, 609-628.
4. Minty, A.J., Alonso, S., Guenet, J.L. and Buckingham, M.E. 1983 *J.Mol. Biol.* 167, 77-101.
5. Ponte, P., Gunning P., Blau, H. and Kedes, L. (1983) *Mol. Cell Biol.* 3, 1783-1791.
6. Vandekerckhove, J. and Weber, K. (1978) *J. Mol. Biol.* 126, 783-802.
7. Vandekerckhove, J. and Weber, K. (1981) *Eur. J. Biochem.* 113, 595-603.
8. Shah, D.M., Hightower, R.C., Meagher, R.B. (1983) *J. of Mol. and App. Genetics* 2, 111-126.
9. Nellen, W. and Gallwitz, D. (1982) *J. Mol. Biol.* 159, 1-18.
10. Fryberg, E.A., Bond, B.J., Hershey, N.D., Mixter, K.S. and Davidson, N. (1981) *Cell* 24, 107-116.
11. Vandekerckhove, J. (1983) 3rd EMBO Workshop on Mol. Cell. Aspects of Myogenesis and Myofibrillogenesis in Cell Cultures of Normal and Diseased Muscle, Zurich, Switzerland pp. 75.
12. Gunning, P., Ponte, P., Kedes, L., Eddy, R. and Shows, T. (1984) *Proc. Natl. Acad. Sci. USA* 81, 1813-1817.
13. Hamada, H., Petrino, M.G. and Kakunaga, T. (1982) *Proc. Natl. Acad. Sci. USA* 79, 5901-5905.
14. Fornwald, J.A., Kuncio, G., Peng, I. and Ordahl, C.P. (1982) *Nucleic Acids Res.* 10, 3861-3876.
15. Zakut, R., Shani, M., Givol, D., Neumah, S., Yaffe, D. and Nudel, U. (1982) *Nature* 298, 857-859.
16. Minty, A.J., Alonso, S., Caravatti, M. and Buckingham, M.E. (1982) *Cell* 30, 185-192.

17. Gunning, P., Ponte, P., Blau, H. and Kedes, L. (1983) *Mol. Cell. Biol.* 3, 1985-1995.
18. Smith, H.O. and Birnstiel, M.L. (1976) *Nucleic Acid Res.* 3, 2387-2389.
19. Messing, J. (1983) *Methods in Enzymol.* 101, 20-78.
20. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Nat. Acad. Sci. USA* 74, 5463-5467.
21. Sanger, F., and Coulson, A.R. (1978) *FEBS Lett.* 87, 107-110.
22. Vandekerckhove J. and Weber, K. (1979) *Differentiation* 14, 123-133.
23. Maxam, A., and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* 74, 560-564.
24. Kost, T.A., Theodorakis, N. and Hughes, S.H. (1983) *Nucleic Acids Res.* 11, 287-8301.
25. Nudel, U., Zakut, R., Shani, M., Neuman, S., Levy, Z. and Yaffe, D. (1983) *Nucleic Acid Res.* 11, 1759-1771.
26. Lewin, B. (1980) *Gene Expression-2*, 2nd Ed., John Wiley and Sons, New York, p. 817.
27. Breathnach, R. and Chambon, P. (1981) *Ann. Rev. Biochem.* 50, 349-383.
28. Ehrlich, M. and Wang, R.Y.H. (1981) *Science* 212, 1350-1357.
29. Mayer, Y., Czosnek, H., Zeelon, P.E., Yaffe, D. and Nudel, H. (1984) *Nucleic Acids Res.* 12, 1087-1100.
30. Ponte, P., Ng, S.-Y., Engel, J., Gunning, P. and Kedes, L. (1984) *Nucleic Acid Research* 12, 1687-1696.
31. Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R., and Tizard (1979) *Cell* 18, 545-558.
32. Bell, G.I., Pictet, R.L., Rutter, W.J., Cordell, B., Tischer, E., and Goodman, H.M. (1980) *Nature* 284, 26-32.
33. Cooke, N.E., Coit, D., Shine, J., Baxter, J.D. and Martial, J.A. (1981) *J. Biol. Chem.* 256, 4007-4016.
34. Page, G.S., Smith, S. and Goodman, H.M. (1981) *Nucleic Acids Res.* 9, 3719-3730.
35. Lu, R.C. and Elzinga, M. (1977) *Biochem.* 16, 5801-5806.