Volume 11 Number 23 1983 Nucleic Acids Research

The nucleotide sequence of the chick cytoplasmic β-actin gene

Thomas A.Kost, Nicholas Theodorakis and Stephen H.Hughes

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

Received 24 August 1983; Revised and Accepted 4 November 1983

ABSTRACT

The nucleotide sequence of the chick β-actin gene was determined. The gene contains 5 introns; 4 interrupt the translated region at codons 41/42, 120/122, 267, 327/328 and a large intron occurs in the 5' untranslated region. The gene has a 97 nucleotide 5'-untranslated region and a 594 nucleotide 3'-untranslated region. A slight heterogeneity in the position of the poly A addition site exists; polyadenylation can occur at either of two positions two nucleotides apart. The gene codes for an mRNA of 1814 or 1816 nucleotides, excluding the poly(A) tail. In contrast to the chick skeletal muscle actin gene the β-actin gene lacks the Cys codon between the initiator ATG and the codon for the N-terminal amino acid of the mature protein. In the 5' flanking DNA, 15 nucleotides downstream from the CCAAT sequence, is a tract of 25 nucleotides that is highly homologous to the sequence found in the same region of the rat β-actin gene.

INTRODUCTION

Actin is an abundant, highly conserved protein that is found in all eukaryotic cells. It is a major component of the cytoskeleton and is involved in cell motility, mitosis and muscle contraction (1). On the basis of amino acid sequence six different forms of actin have been identified in vertebrates (2). These include two cytoplasmic actins β and γ , and four muscle specific α-actins. The muscle specific actins are tissue specific and include a skeletal muscle actin, cardiac muscle actin and two smooth muscle actin forms. The muscle specific actins are more closely related to each other in primary structure than they are to the cytoplasmic β and γ actins. The vertebrate nonmuscle β and γ-actins are more closely related to the actins found in lower eukaryotes than are the vertebrate muscle actins.

As indicated by genomic blots, the number of actin genes in

© IRL Press Limited, Oxford, England.

higher organisms varies widely. In the chick, 4-7 actin genes can be detected (3). Human DNA appears to contain 20-30 genes (4,5), mouse greater than 20 (6) and rat 12 or more (7). It has not yet been determined whether all of these genes encode different functional actins or if some are actin-like pseudogenes. Recently Moos and Gallwitz (8) described the isolation and sequence of two human β -actin like pseudogenes. The number of actin genes found in lower eucaryotes is also highly variable. Drosophila contains six genes (9), yeast-1 (10-11), Dictyostelium-17 (12) and sea-urchin-11 (13).

Actin gene expression is tissue specific and developmentally regulated (14,15, Kost et al., unpublished results). By studying the structural organization of the actin genes, one can begin to answer questions concerning the evolution of this gene family and develop tools for the analysis of the control of expression of these genes during development. In this report we present the complete nucleotide sequence of the chicken β-actin gene and compare its structure to that of other actin genes whose sequence is known.

MATERIALS AND METHODS

Materials

Restriction endonucleases, DNA polymerase I (Klenow fragment) and T4 DNA ligase were obtained from New England Biolabs; T4 polynucleotide kinase from P-L Biochemicals; calf intestinal alkaline phosphatase from Boehringer Mannheim; linkers and oligo dT cellulose from Collaborative Research; AMV reverse transcriptase from Life Sciences Inc., and ^{32}P-radio-nucleotides from Amersham Radiochemicals.

Screening the Genomic Library

A chicken genomic DNA library prepared by Dodgson et al. (16) was screened for the β-actin gene as described previously (17). For this study the phage λβAct-1 containing 8.9 kb of chicken genomic DNA was used. Phage DNA was prepared as described by Yamamoto et al. (18).

Subcloning Fragments into PBR322

Plasmid DNA was prepared using the procedure described by

Holmes and Quigley (19), followed by cesium chloride-ethidium bromide density gradient centrifugation. λβAct-1 DNA was digested with EcoRI and the 8.9 kb fragment containing the β-actin gene and flanking DNA was isolated by electrophoresis through a 0.8% agarose gel. The 8.9 kb fragment was electroeluted, purified on a DEAE-sephacel column and ligated to EcoRI digested PBR322. The ligation mixture was used to transform E.coli DH1 as described by Hanahan (20). The recombinant plasmid containing the β-actin gene is referred to as pβAct-1. Two subclones of pβAct-1 were constructed by digestion of the plasmid with EcoRI and HindIII followed by ligation of the purified fragment into PBR322 from which the EcoRI-HindIII fragment had been removed. The resulting plasmids, pβAct-2 and pβAct-3 contain the 5' end and the 3' end of the β-actin gene, respectively.

Bal31 Digestion and Linker Insertion

For sequence determination of the β-actin gene a series of plasmids containing overlapping fragments of the gene were constructed. Ten μg of pβAct-3 were digested with HindIII, extracted with phenol:chloroform (1:1) and precipitated with two volumes of ethanol. The DNA was dissolved in 80 μl of 0.2M NaCl, 12mM $MgCl_2$, 12mM $CaCl_2$, 20mM Tris·HCl, pH 8.0, 1mM EDTA containing 0.3 units of Bal31. The sample was placed at 37°C and at various times aliquots were removed and the digestion terminated by the addition of EGTA to a final concentration of 20mM. The samples were extracted with a 1:1 mixture of phenol: chloroform followed by precipitation with 2 volumes of ethanol. Recessed 3' termini were filled in using dNTPs and DNA polymerase I (Klenow fragment) followed by ligation to HindIII linkers. E.coli DH1 was transformed with the plasmids and the resulting colonies screened for those containing plasmids smaller than the parent pβAct-3. To obtain a full collection of plasmids some of the smaller derivatives which were isolated were linearized and treated with Bal31. In this manner a collection of plasmids was obtained each having an insert 100-400 bases shorter than the next largest plasmid. A collection of plasmids covering the 5' end of the gene was constructed in a similar manner by digestion of pβAct-2 with XhoI followed by limited Bal31 digestion and

ligation of XhoI linkers. This strategy obviates the need for a detailed restriction map prior to sequence analysis.

DNA sequencing

Plasmids were digested with HindIII or XhoI and were end-labeled using E.coli DNA polymerase I (Klenow fragment). The plasmids were then digested with EcoRI or HindIII and the single end-labeled fragment isolated by agarose gel electrophoresis. The fragments were purified on DEAE-sephacel columns and sequenced by the chemical method described by Maxam and Gilbert (21). The cleaved fragments were fractionated on 40 or 80cm 0.3mm urea-acrylamide gels. The gels were lifted onto Whatman 3MM filter paper and dried prior to autoradiography. In regions having ambiguities the complementary strand was sequenced.

Primer Extension Analysis

Chick fibroblast poly(A) RNA was isolated by guanidine thiocyanate extraction (22) and oligo dT cellulose chromatography (23). pβAct-3 was digested with BglII and 5' end labeled with polynucleotide kinase and ^{32}P-ATP. The plasmid was then digested with HaeIII and fractionated by electrophoresis through a 5% acrylamide gel. A 108bp fragment spanning positions 464-571 (Figure 2) was isolated. One pmole of the labeled fragment was denatured in 80 μl of 98% formamide containing 2 μl of 0.5M EDTA by heating at 80°C for 10 min. Following denaturation 5 μg of poly(A) in 20 μl of 200mM Pipes (pH 6.4), 2.0M NaCl, 5mM EDTA was added and the mixture incubated at 47°C for 16 hr. The nucleic acids were precipitated by the addition of 70 μl of 5M NH_4 acetate and 500 μl of ethanol. The precipitate was collected by centrifugation, redissolved in 0.3M Na acetate and precipitated with ethanol. The dried precipitate was suspended in 45 μl of 50mM Tris·Cl, pH 8.3, 20mM KCl, 10mM $MgCl_2$, 5mM DTT, 1.25 μl each of 20mM dNTPs and 20 units of reverse transcriptase. Incubation was at 42°C for 1 hr. The nucleic acid was extracted with an equal volume of phenol:chloroform, precipitated with two volumes of ethanol and fractionated on an 80cm 5% sequencing gel and autoradiographed.

RESULTS AND DISCUSSION

Our laboratory previously reported the isolation of the chick β-actin gene from a Charon 4A library of chicken DNA (17). The 8.9 kb insert containing the structural gene and flanking DNA was subcloned into the EcoRI site of pBR322 (pβAct-1, Fig. 1A). The two EcoRI-HindIII fragments of pβAct-1 were also subcloned into pBR322. The resulting recombinant plasmids pβAct-2 and pβAct-3 contain the 5' and 3' regions of the β-actin gene, respectively. From these clones a series of Bal31 digested derivatives were prepared and used to determine the nucleotide sequence of the gene and flanking DNA as described in Materials and Methods.

A detailed restriction map of the 5046 nucleotide segment that was sequenced is shown in Fig. 1B. Enzymes which only cut at one site in the gene include HindIII, BglII, StuI, MstI, NcoI

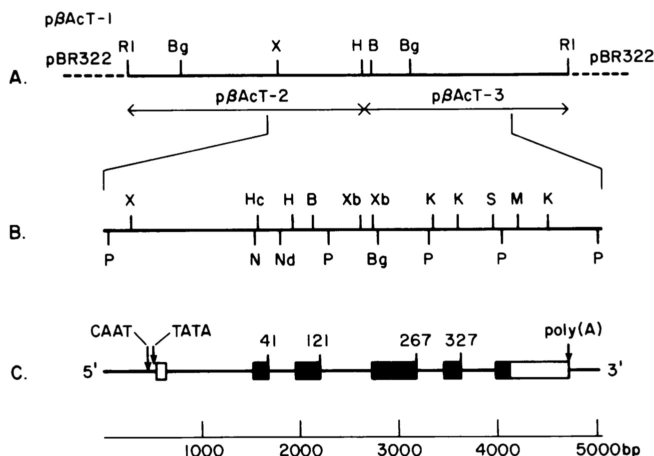

Figure 1. Structural map of the chick β-actin gene.
(A) Restriction map of the 8.9 kb EcoRI fragment of pβ Act-1 containing the β-actin gene. pβ Act-2 and pβ Act-3 are the HindIII-EcoRI fragments that were subcloned from p Act-1.
(B) Detailed restriction map of the β-actin structural gene and flanking DNA sequence. Restriction sites shown in (A) and (B) are: BglII (B); BglI (Bg); EcoRI (RI); HincII (Hc); HindIII (H); Kpn (K); MstI (M); NcoI (N): NdeI (Nd); PstI (P); StuI (S); XbaI (Xb); XhoI (X).
(C) Schematic representation of the structure of the β-actin gene. Solid bars: coding region; open bars: untranslated region; solid line: introns and flanking DNA. The numbers refer to codon positions.

A

```
CACCGGTGTTATTGCTGCTCGGTGCGTGCATGCACATCAGTGTCGCTGCAGCTCAGTGCATGCACGCTCATTG
                                                -1500
CCCATCGCTATCCCTGCCTCTCCTGCTGGCGCTCCCCGGGAGGTGACTTCAAGGGGACCGCAGGACCACCTCG
                    -1450                                                -1400
GGGGTGGGGGGAGGGCTGCACACGCGGACCCCGCTCCCCCTCCCCAACAAAGCACTGTGGAATCAAAAAGGGG
                                                     -1350
GGAGGGGGGATGGAGGGGCGCGTCACACCCCCGCCCCACACCCTCACCTCGAGGTGAGCCCCACGTTCTGCTT
                         -1300
CACTCTCCCCATCTCCCCCCCCTCCCCACCCCCAATTTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCG
1250                                                      -1200
ATGGGGGCGGGGGGGGGGGGGCGCGCCCAGGCGGGGCGGGGCGGGGCGAGGGGCGGGGCGGGGCGAGGCGG
                              -1150
AGAGGTGCGGCGGCAGCCAATCAGAGCGGCGCGCTCCGAAAGTTTCCTTTTATGGCGAGGCGGCGGCGGCGGC
   -1100                                                        -1050
GGCCCTATAAAAAGCGAAGCGCGCGGCGGGCGGGAGTCGCTGCGTTGCCTTCGCCCCGTGCCCCGCTCCGCGC
                                  -1000
CGCCTCGCGCCGCCCGCCCCGGCTCTGACTGACCGCGTTACTCCCACAG / GTGAGCGGGCGGGACGGCCCT
       -950                                                   -900
TCTCCTCCGGGCTGTAATTAGCGCTTGGTTTAATGACGGCTCGTTTCTTTTCTGTGGCTGCGTGAAAGCCTTA
                                          -850
AAGGGCTCCGGGAGGGCCCTTTGTGCGGGGGGGAGCGGCTCGGGGGTGCGTGCGTGTGTGTGTGCGTGGGGA
                 -800                                                  -750
GCGCCCGCGTGCGGCCCGCGCTGCCCGGCGGCTGTGAGCGCTGCGGGCGCGGCGCGGGGCTTTGTGCGCTCCGC
                                              -700
GTGTGCGCGAGGGGAGCGCGGCCGGGGGCGGTGCCCCGCGGTGCGGGGGGGCTGCGAGGGGAACAAAGGCTGC
                       -650                                              -600
GTGCGGGGTGTGTGCGTGGGGGGGTGAGCAGGGGGTGTGGGCGCGGCGGTCGGGCTGTAACCCCCCCCTGCAC
                                                     -550
CCCCCTCCCCGAGTTGCTGAGCACGGCCCGGCTTCGGGTGCGGGGCTCCGTGCGGGGCGTGGCGCGGGGCTCG
                              -500
CCGTGCCGGGCGGGGGGTGGCGGCAGGTGGGGGTGCCGGGCGGGGCGGGGCCGCCTCGGGCCGGGGAGGGCTC
450                                                        -400
GGGGGAGGGGCGCGGCGGCCCCGGAGCGCCGGCGGCTGTCGAGGCGCGGCGAGCCGCAGCCATTGCCTTTTAT
                                     -350
GGTAATCGTGCGAGAGGGCGCAGGGACTTCCTTTGTCCCAAATCTGGCGGAGCCGAAATCTGGGAGGCGCCGC
   -300                                                          -250
CGCACCCCCTCTAGCGGGCGCGGGCGAAGCGGTGCGGCGCCGGCAGGAAGGAAATGGCGGGGAGGGCCTTCG
                                          -200
TGCGTCGCCGCGCCGCCGTCCCCTTCTCCATCTCCAGCCTCGGGGCTGCCGCAGGGGGACGGCTGCCTTCGGG
        -150                                                          -100
GGGGACGGGGCAGGGCGGGGTTCGGCTTCTGGCGTGTGACCGGCGGGGTTTATATCTTCCCTTCTCTGTTCCT
                                               -50
CCGCAG / CCAGCC
              -1
```

B

```
                                                 10
Met Asp Asp Asp Ile Ala Ala Leu Val Val Asp Asn Gly Ser Gly Met Cys Lys
ATG GAT GAT GAT ATT GCT GCG CTC GTT GTT GAC AAT GGC TCC GGT ATG TGC AAG
1                                                                    50
    *20                                         *30
Ala Gly Phe Ala Gly Asp Asp Ala Pro Arg Ala Val Phe Pro Ser Ile Val Gly
GCC GGT TTC GCC GGG GAC GAT GCC CCC CGT GCT GTG TTC CCA TCT ATC GTG GGT
                                                                100
            *40
Arg Pro Arg His Gln
CGC CCC AGA CAT CAG / GTACGGCAACGGCTGCGGTGCTGCATCCGGACGATAGGGAGGAGCGTTC
                                                         150
GGTGGGAGTGGTTCTGCTGAGGGCAGCGAGCCTCAAGAAGCCCTTTTATTTTCATATGTGACATAAGGAGT
                                 200
TTCTGTCCTCAACAGTAATGGGTTGAGTACGCAGCCTCCGCGGAGCATCCTGTGTTGGAGCAGTTGCTCAG
      250                                                        300
```

```
TCCCTTCCCTGACAGAGTAGGGACAGTGGGGGGTTGGTAATTAAAGGTGTGGCACT[illegible]GGAAAGAAAACTTT
                                                    350
                                                                                Gly Val
ATTTGTCCGTTGGAGTTGATCCTTCAGTTAAACTTAAAGCTTTTTCTTTGTTG[illegible]CAG / GGT GTG
                  400
                                       *50                                          *60
Met Val Gly Met Gly Gln Lys Asp Ser Tyr Val Gly Asp Glu Ala Gln Ser Lys
ATG GTT GGT ATG GGC CAG AAA GAC AGC TAC GTT GGT GAT GAA GCC CAG AGC AAA
450                                                                                   500
                                                  *70
Arg Gly Ile Leu Thr Leu Lys Tyr Pro Ile Glu His Gly Ile Val Thr Asn Trp
AGA GGT ATC CTG ACC CTG AAG TAC CCC ATT GAA CAC GGT ATT GTC ACC AAC TGG
                                                                               550
*80                                                        *90
Asp Asp Met Glu Lys Ile Trp His His Thr Phe Tyr Asn Glu Leu Arg Val Ala
GAT GAT ATG GAG AAG ATC TGG CAC CAC ACT TTC TAC AAT GAG CTG AGA GTA GCC
                                                                          600
        *100                                                      *110
Pro Glu Glu His Pro Val Leu Leu Thr Glu Ala Pro Leu Asn Pro Lys Ala Asn
CCT GAG GAG CAC CCT GTG CTG CTC ACA GAG GCT CCC CTG AAC CCC AAA GCC AAC
                                                                   650
                  *120
Arg Glu Lys Met Thr Gln
AGA GAG AAG ATG ACA CAG / GTGTGTAAAACTCTTTGAGCCTAGAGCTACAGCAGGTACTGCTGC
                                                             700
AGGCCAGCTTTCCTCTTCCATGTTGTCCTCTATCTGCCTTCACCACTTCTCCCTTTTTGCCATCTTTACAG
                                     750
GGTTTTCCTTTCCTGACCTGAGTCTCCTCTTTGCTGGACCTTGACAGGTTTTGTTTGCTCTAAGCTGGCTT
800                                                                 850
TCTCTGAGACTGAACTAGCAACTTGTCTAACTGCTGTTTCTGACTAGACACTAATCCATTTACCATCTTTG
                                              900
AGTGACTATACCGTAGTTTCTTGGCCTTTGTTTCGTTCCTTGCTGTGTCATGTGGATGGGCCTGTGGTAGC
           950                                                                 1000
AGTGTCCTTGCTCAGGCTCTGACTTGACAGGTGTGGGTGGAGACCAAGTGAGATAAGTGCTCATAGCTCCT
                                                       1050
TCTAGAGGAGCAGAGAAGCCTCTTATCTAGGGACAAATGTAAACCACATGCATGGACCGATGTAGGTGTGC
                  1100                                                                                 1150
                                                                           Ile Met Phe Glu
AGTGAGAACCTTGGTTTATTCTGCTTGACTCAGTGCTTTTTTTATCTTTCTAG / ATC ATG TTT GAG
                                                                   1200
                 *130                                                        *140
Thr Phe Asn Thr Pro Ala Met Tyr Val Ala Ile Gln Ala Val Leu Ser Leu Tyr
ACC TTC AAC ACC CCA GCC ATG TAT GTA GCC ATC CAG GCT GTG CTG TCC CTG TAT
                                                           1250
                                     *150                                          *160
Ala Ser Gly Arg Thr Thr Gly Ile Val Met Asp Ser Gly Asp Gly Val Thr His
GCC TCT GGT CGT ACC ACT GGT ATT GTG ATG GAC TCT GGT GAT GGT GTT ACC CAC
                                                     1300
                                               *170
Thr Val Pro Ile Tyr Glu Gly Tyr Ala Leu Pro His Ala Ile Leu Arg Leu Asp
ACT GTG CCC ATC TAT GAA GGC TAC GCC CTC CCC CAT GCC ATC CTC CGT CTG GAT
                                              1350
180                                                         *190
Leu Ala Glv Arg Asp Leu Thr Asp Tyr Leu Met Lys Ile Leu Thr Glu Arg Gly
CTG GCT GG  CGT GAC CTG ACG GAC TAC CTC ATG AAG ATC CTG ACA GAG AGA GGC
                                     1400
        *20C                                                        *210
Tyr Ser Phe Thr Thr Thr Ala Glu Arg Glu Ile Val Arg Asp Ile Lys Glu Lys
TAC AGC TTC ACC ACC ACA GCC GAG AGA GAA ATT GTG CGT GAC ATC AAG GAG AAG
                            1450
                  *220                                                        *230
Leu Cys Tyr Val Ala Leu Asp Phe Glu Gln Glu Met Ala Thr Ala Ala Ser Ser
CTG TGC TAC GTC GCA CTG GAT TTC GAG CAG GAG ATG GCC ACA GCT GCC TCT AGC
                  1500
```

```
  *234A                   *240                                        *250
Ser Ser Leu Glu Lys Ser Tyr Glu Leu Pro Asp Gly Gln Val Ile Thr Ile Gly
TCT TCC CTA GAG AAG AGC TAT GAA CTC CCT GAT GGT CAG GTC ATC ACC ATT GGC
        1550
                                  *260
Asn Glu Arg Phe Arg Cys Pro Glu Ala Leu Phe Gln Pro Ser Phe Leu G
AAT GAG AGG TTC AGG TGC CCC GAG GCC CTC TTC CAG CCA TCT TTC TTG G/ GTAAG
  1600                                                                   1650
TCTGACTCCTTTGAGGAGATTCCATGCCCCTTCTGCCTAGAGCGACTTAGAGACTGGCCACAATATTGCTGG
                                                          1700
CCCTTAGCAGGTCTGTCCCATTCCACTGCACTCTTTTTCTTCCTCAGTTTGAGGATGATTGAGGTACCTA
                          1750
GGTCAGTAGAACAGGAGGCTGAGCCTGTCAGCCATGCCTTTGTCTGTGTTGTTTCCCACTATGGGGTGTTT
     1800                                                 1850
GATTCTCTTGTTTTCCCATAGCAGTAGGTCCTCTGTCTTCCTTGTCCCTGAACCATTTACTGACTTGACAC
                                   1900
                              *270
                        ly Met Glu Ser Cys Gly Ile His Glu Thr Thr Phe Asn
TGCTTTTTGCTTGCAG / GT ATG GAG TCC TGT GGT ATC CAT GAA ACT ACC TTC AAC
              1950
280                                     *290
Ser Ile Met Lys Cys Asp Val Asp Ile Arg Lys Asp Leu Tyr Ala Asn Thr Val
TCC ATC ATG AAG TGT GAT GTG GAT ATC CGT AAG GAT CTG TAT GCC AAC ACA GTG
              2000
     *300                                     *310
Leu Ser Gly Gly Thr Thr Met Tyr Pro Gly Ile Ala Asp Arg Met Gln Lys Glu
CTG TCT GGT GGT ACC ACA ATG TAC CCT GGC ATT GCT GAC AGG ATG CAG AAG GAG
    2050
             *320
Ile Thr Ala Leu Ala Pro Ser Thr Met Lys Ile Lys
ATC ACA GCC CTG GCA CCT AGC ACA ATG AAA ATC AAG / GTAGGCTGGAGCCCTAGGCTG
 2100                                                              2150
CTCTTAGCACACCTCTTAATGCTGAGTGGGTGGAGGAGGCTGAGCTAAACTACTACAGAAGATGGAAACTG
                                                          2200
CTTTGTCTTAGTGGTGGAGCCTTATGGCCTCGCTCTGGGCAGGGTGGCAGCCACTCAGCCATGTGAGTAAT

GAGATTAAGTGCTGGCTGTGAGTTGGGGTGGAGGGGGAGGGTGCTGTGTGTAGGGCTGTGCCCTTAAGCCT
  2300                                                    2350
GCTCAGACTCTGGAGTGCCTTGGCCAAGGCCTGATAAAGGAATTAGGTAGCCGGAGCGCCACAGGGCTGCGT
                                2400
                                                              *330
                                                           Ile Ile Ala Pro
GTGGGTTGTGTACCATCACACTGCATCTGACCAGTTGTTCTCTCCTGCAG / ATC ATT GCC CCA
          2450                                                         2500
                                  *340
Pro Glu Arg Lys Tyr Ser Val Trp Ile Gly Gly Ser Ile Leu Ala Ser Leu Ser
CCT GAG CGC AGG TAC TCT GTC TGG ATT GGA GGC TCT ATC CTG GCC TCC CTG TCC
                                                                    2550
350                                     *360
Thr Phe Gln Gln Met Trp Ile Ser Lys Gln Glu Tyr Asp Glu Ser Gly Pro Ser
ACC TTC CAG CAG ATG TGG ATC AGC AAG CAG GAG TAC GAT GAA TCC GGA CCC TCC
                                                            2600
     *370
Ile Val His Arg Lys Cys Phe End
ATT GTC CAC CGC AAA TGC TTC TAAACCGGACTGTTACCAACACCCACACCCCTGTGATGAAACA
                                                         2650
AAACCCATAAATGCGCATAAAACAAGACGAGATTGGCATGGCTTTATTTGTTTTTCTTTTGGCGCTTGAC
                          2700
TCAGGATTAAAAAACTGGAATGGTGAAGGTGTCAGCAGCAGTCTTAAAATGAAACATGTTGGAGCGAACGC
     2750                                                 2800
CCCCAAAGTTCTACAATGCATCTGAGGACTTTGATTGTACATTTGTTTCTTTTTTAATAGTCATTCCAAAT
                                    2850
ATTGTTATAATGCATTGTTACAGGAAGTTACTCGCCCTCTGTGAAGGCAACAGCCCAGCTGGGAGGAGCCGG
              2900
TACCAATTACTGGTGTTAGATGATAATTGCTTGTCTGTAAATTATGTAACCCAACAAGTGTCTTTTTGTAT
                                                  3000
```

```
CTTCCGCCTTAAAAACAAAACACACTTGATCCTTTTTGGTTTGTCAAGCAAGCGGGCTGTGTTCCCCAGTG
                          3050
ATAGATGTGAATGAAGGCTTTACAGTCCCCCACAGTCTAGGAGTAAAGTGCCAGTATGTGGGGAGGGAGG
3100                                                      3150
GGCTACCTGTACACTGACTTAAGACCAGTTCAAATAAAAGTGCACACAATAGAGGCTTGACTGGTGTTGGT
                                   3200
TTTTATTTCTGTGCTGCGCTGCTTGGCCGTTGGTAGCTGTTCTCATCTAGCCTTGCCAGCCTGTGTGGGTC
        3250                                                    3300
AGCTATCTGCATGGCTGCGTGCTGGTGCTGTCTGGTGCAGAGGTTGGATAAACCGTGATGATATTTCAGC
                                           3350
AAGTGGGAGTTGGCTCTGATTCCATCCTGAGCTGCCATCAGTGTGTTCTGAAGGAAGCTGTTGGATGAGGG
                  3400                                                3450
TGGGCTGAGTGCTGGGGGACAGCTGGGCTCAGTGGGACTGCAGCTGTGCT
                                                  3500
```

Figure 2. Nucleotide sequence of the chick β -actin gene. (A) 5' untranslated region and flanking DNA sequence; (B) coding, 3' untranslated region and flanking DNA sequences. The nucleotides are numbered relative to the A of the Met codon used to initiate transcription of the gene. Nucleotides 5' of the point are designated by negative numbers. The CCAAT, TATA box, cap site and polyadenylation signal are underscored. Slash marks represent exon-intron junctions. Translated nucleotides are separated into codons and numbered as proposed by Lu and Elzinga (24) with codon numbers preceded by an *.

and NdeI. A schematic representation of the β-actin gene is shown in Fig. 1C. The translated region of the gene is interrupted by four introns and a large intron is present in the 5' untranslated region.

<u>The Translated Region of the β-Actin Gene</u>

The complete nucleotide sequence of the β-actin gene with 5' and 3' flanking DNA is shown in Figs. 2A, B. The nucleotide sequence codes for a protein which is identical to the bovine β-actin (2), to human fibroblast β-actin (25), and to the protein sequence predicted by the nucleotide sequence of the rat β-actin gene (26). This shows the strong evolutionary conservation of the β-actins.

The coding sequence of the chick skeletal muscle α-actin gene (27) as well as the rat skeletal muscle α-actin gene (28) begins with a met-cys dipeptide which is absent from the mature protein. This dipeptide is also found in all six drosophila actin genes (9), a human cardiac muscle actin gene (29) and in sea urchin actin genes (30). In contrast the cys codon is absent in the chick β-actin gene, and the rat β-actin gene (26). The chick β-actin gene codes for a mature protein

Table 1. SUMMARY OF -ACTIN GENE ORGANIZATION

Exon	Intron	Positions	Size (nt)	Coding region of mRNA
1		-1000-(-910)	91	91 of 97 nt of 5' untrans.
	1	-909-(-7)	903	
2		-6-123	129	a.a.1-41+6 nt of 5' untrans.
	2	124-444	320	
3		443-683	240	a.a.42-121
	3	684-1207	524	
4		1208-1646	439	a.a.122-266
	4	1647-1952	306	
5		1953-2133	182	a.a.267-327
	5	2134-2489	355	
6		2490-3224	735	a.a.328-374+594 nt of 3' untrans.
		Total	4224	

that is one amino acid shorter than the vertebrate skeletal muscle α-actins (26,27).

Nudel et al. (26) have previously summarized the frequency of codon usage in the various actin genes. The overall pattern of codon usage in the rat and chicken skeletal muscle α-actin genes, a human cardiac actin gene, the rat β-actin gene and chick β-actin gene is very similar.

The coding region of the β-actin gene is interrupted by four introns at the codons specifying amino acids 41/42, 121/122, 267 and 327/328 (Fig. 1C). The size and positions of the introns are summarized in Table 1. The sequence of the splice sites agrees with the consensus sequence for splice sites compiled by Mount (31). The position of the exon-intron boundaries are identical to those of the rat β-actin gene, however, the introns in the coding region of the chick gene are larger than the corresponding introns in the rat gene. No sequence homology is found between the introns in the rat and chick β-actin genes or the chick β and chick skeletal muscle α-actin gene.

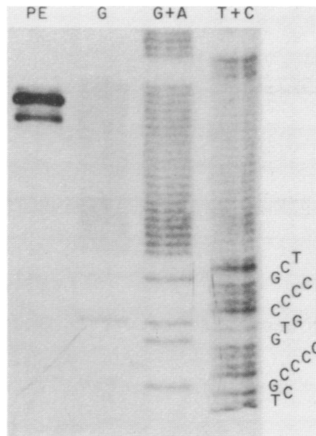

Figure 3. Primer extension mapping of the 5' cap site of the β-actin gene.
The 5'-end labeled 108 nucleotide BglII - HaeIII fragment spanning positions 464-571 (Figure 2) was hybridized to chick fibroblast poly A^+ RNA and extended using reverse transcriptase as described in Materials and Methods. For size markers a sequencing ladder was prepared from a nearly full length chick β-actin cDNA by labeling at the same BglII site as the primer fragment.

The chick and the rat skeletal muscle α-actin genes have five introns in the translated region of the gene (26,27). The introns at codons 41/42, 267 and 327/328 are shared by the β-actin and skeletal muscle α-actin genes. The skeletal genes lack the intron between codons 121/122 and have two additional introns at codons 150 and 204. A summary of intron positions in actin genes from various organisms has been previously reported by Fornwald et al. (27).

The 5' Untranslated Region

The initial electronmicrographs of R-loops formed between the β-actin gene and mRNA from chick embryo fibroblasts did not indicate the presence of an intron in the 5' untranslated region (17). However sequence analysis of the 5' end of a nearly full length chick β-actin cDNA clone (data not shown) revealed that an intron interrupted the 5' untranslated region of the β-actin gene six nucleotides upstream of the initiator ATG (Fig. 2A). The 5' terminus of the intron was determined by aligning the

5'—> 3'

Chick β-actin

GCGGCGGCAGCCAATCAGAGCGGCGCGCTCCGAAAGTTTCCTTTTATGGCGAGGCGGCGGCGGCGGCGGCCCTAT

Rat β-actin

CGAGCCGGAGCCAATCAGCGCCCGCCGTTCCGAAATTGCCTTTTATGGCTCGAGTGGCCGCTGTGGCGTCCTATA

Chick skeletal muscle α-actin

GCCCGACACCCAAATATGGCGCGGCCGGGGTCGCATTCCTGTCGCGGGCCGGGCGGTGCTCCCGTCGATAAAA

Rat skeletal muscle α-actin

GCCCAACACCCAAATATGGCTTGGAAGGGCAACAACATTCCTTCGGGCGGTGTGGAGAGCTCAGGACTATATAA

Figure 4. Sequence homologies of chick and rat β and skeletal muscle α-actin genes in the CAAT-TATA box region. Regions of homology are underscored. Rat β-actin gene (26), chick skeletal muscle α-actin gene (27), rat skeletal muscle α-actin gene (28).

genomic sequence with the sequence of the β-actin cDNA sequence. A match in the genomic sequence with that of the cDNA begins at position -910 and continues to position -981 which marks the 5' terminus of the cDNA clone. The intron is therefore 903 nucleotides long. The cap site was determined by primer extension of a probe prepared from within the third exon (Fig. 3). The fully extended product corresponds to the G at position -999 in Fig. 2A. A strong band is also seen that corresponds to the A at position 997 and may represent a secondary transcription initiation site. A Goldberg-Hogness TATA box sequence (32,33) TATAAAA, is located 27 nucleotides upstream of the cap site and a CCAAT sequence (34,35) is located 89 nucleotides upstream of the cap site. The relative position of these putative regulatory sequences corresponds well with the position of similar sequences found upstream of the 5' termini of other eukaryotic genes (33).

A comparison of the CCAAT-TATA box region of the chick β-actin gene with the rat β-actin gene is shown in Fig. 4. A block of 25 nucleotides beginning 15 nucleotides downstream from

the C of the CCAAT sequence is highly conserved in this region. Despite the presence of homology near the CCAAT sequence the 5'-untranslated regions of the chick and rat β-actin genes show no homology. The strong conservation of the sequence block near the CCAAT sequence suggests that it may function in the control of transcriptional initiation of the β-actin gene. The same region is not conserved in the chick or rat skeletal muscle α-actin genes (Fig. 4). Fig. 4 also shows a block of 20 nucleotides surrounding the CCAAT sequence that is conserved (except for one mismatch) in the rat and chick skeletal muscle α-actin genes (26,27) and not in the β-actin genes. As proposed by Ordahl and Cooper (36) this sequence may also represent a control region for the tissue specific activation of these genes. Future experiments which utilize the β-actin and skeletal muscle α-actin genes mutagenized in these regions should be useful in determining the actual function of these conserved sequences.

The 3' Untranslated Region

The size of the 3' untranslated region was determined by comparing the nucleotide sequence of two independent cDNA clones both of which contain the complete 3' untranslated region of the β-actin gene, with the sequence of the 3' end of the β-actin gene. The polyA tract of one of the cDNA clones begins at the G corresponding to position 3216 (Fig. 2) of the β-actin gene. The polyA tract of the other cDNA clone begins at the G corresponding to position 3214 of the β-actin gene (D. Cleveland, personal communication). Thus it appears that the mRNA for β-actin can terminate at two possible sites two nucleotides apart preceeding the poly(A) tail. A potential polyadenylation signal AAATAAAA (37) is located 20 or 22 nucleotides upstream from the poly(A) addition sites.

There are a number of sequence blocks in the 3' untranslated region which are conserved between the chick β-actin gene and the rat β-actin gene (26). There are also conserved regions in the 3' untranslated regions of the chick and rat skeletal muscle α-actin genes, however, no conservation of sequence in this region is observed between the β-actin and skeletal muscle α-actin genes (26,27). Further experiments are necessary to

determine if these conserved sequences have functional significance.

ACKNOWLEDGEMENTS

This work was supported by a grant from the National Institutes of Health. T.A.K. was the recipient of an NRSA Fellowship and a Leukemia Society of America Special Fellowship. N.T. was supported by the Cold Spring Harbor Laboratory undergraduate research program. We also thank Don Cleveland for providing the β-actin cDNA clone.

REFERENCES

1. Pollard, T.D. and Weihing, R.R. (1974), Crit. Rev. Biochem 2: 1-65.
2. Vandekerckhove, J. and Weber, K. (1979), Differentiation 14: 123-133.
3. Cleveland, D.W., Lopata, M.A., MacDonald, R.J., Cowan, N.J., Rutter, W.J. and Kirschner, M.W. (1980), Cell 20: 95-105.
4. Moos, M. and Gallwitz, D. (1982), Nucl. Acids Res. 10: 7843-7849.
5. Engel, J., Gunning, P., and Kedes, L. (1982), Mol. Cell. Biol. 2: 674-684.
6. Minty, A.J., Alonso, S., Guenet, J.L. and Buckingham, M.E. (1983), J. Mol. Biol. 167: 77-101.
7. Nudel, U., Zakut, R., Katcoff, D., Carmon, Y., Czosnek, H., Shani, M. and Yaffe, D. (1982) In Muscle Development, Pearson, M.L. and Epstein, E.F., Eds., pp. 177-188, Cold Spring Harbor Laboratory, New York.
8. Moos, M. and Gallwitz, D. (1983), The EMBO Journal 2: 757-761.
9. Fyrberg, E.A., Bond, B.J., Hershey, N.D., Mixter, K.S. and Davidson, N. (1981), Cell 24: 107-116.
10. Gallwitz, D. and Sures, I. (1980), Proc. Natl. Acad. Sci. USA 77: 2546-2550.
11. Ng, R. and Abelson, J. (1980), Proc. Natl. Acad. Sci. USA 77: 3912-3916.
12. McKeown, M. and Firtel, R.A. (1981), Cell 24: 799-807.
13. Scheller, R.H., McAllister, L.B., Crain, Jr., W.R., Durica, D.S., Posakony, J.W., Thomas, T.L., Britten, R.J. and Davidson, E.H. (1981), Mol. Cell. Biol. 1: 609-628.
14. Ordahl, C.P., Tilghman, S.M., Ovitt, C., Fornwald, J., Largen, M.T. (1980), Nucl. Acids. Res. 8: 4989-5005.
15. Shani, M., Zevin-Sonkin, D., Carmon, Y., Czosnek, H., Nudel, U. and Yaffe, D. 1982 In Muscle Development, Pearson, M.L. and Epstein, H.F., Eds., pp. 189-200, Cold Spring Harbor Laboratory, 189-200.
16. Dodgson, J.B., Strommer, J. and Engel, J.D. (1979), Cell 17: 879-887.
17. Hughes, S.H., Sorge, J., Thomas, G.P., Feramisco, J.,

Chow, L. and Garrels, J. (1982), Cold Spring Harbor Symp. Quant. Biol. 46: 507-512.
18. Yamamoto, K.R., Alberts, B.M., Benzinger, R., Lawhorne, L. and Treiber, G. (1970), Virology 40: 734-744.
19. Holmes, D.S. and Quigley, M. (1981), Anal. Biochem. 114: 193-197.
20. Hanahan, D. (1983), J. Mol. Biol. (in press).
21. Maxam, A.M. and Gilbert, W. (1980), Methods in Enzymology 65: 499-560.
22. Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979), Biochemistry 24: 5294-5299.
23. Aviv, H. and Leder, P. (1972), Proc. Natl. Acad. Sci. USA 69: 1408-1412.
24. Lu, R.C. and Elzinga, M. (1977), Biochem. 16: 5801-5806.
25. Vandekerckhove, J., Leavitt, J., Kakunaga, T. and Weber, K. (1980), Cell 22: 893-899.
26. Nudel, U., Zakut, R., Shani, M., Neuman, S., Levy, Z. and Yaffe, D. (1983), Nucl. Acids Res. 11: 1759-1771.
27. Fornwald, J.A., Kuncio, G., Ping, I. and Ordahl, C.P. (1982), Nucl. Acids Res. 10: 3861-3875.
28. Zakut, R., Shani, M., Givol, D., Neuman, S., Yaffe, D. and Nudel, U. (1982), Nature 298: 857-859.
29. Hamada, H., Petrino, M.G. and Katunaga, T. (1982), Proc. Natl. Acad. Sci. USA 79: 5901-5905.
30. Cooper, A.D. and Crain, W.R. Jr. (1982), Nucl. Acids Res. 10: 4081-4092.
31. Mount, S.M. (1982), Nucl. Acids Res. 10: 459-471.
32. Goldberg, M.L. (1979), Ph.D. thesis, Stanford University.
33. Breathnach, R., and Chambon, P. (1981), Ann. Rev. Biochem. 50: 349-383.
34. Benoist, C., O'Hare, K. and Chambon, P. (1980), Nucl. Acids Res. 8: 127-142.
35. Efstratiadis, A., Posabony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Bardle, F.E., Shoulders, C.G., Proudfoot, N.J. (1980), Cell 21: 653-668.
36. Ordahl, C.P. and Cooper, T.A. (1983), Nature 303: 348-349.
37. Proudfoot, N.J. and Brownlee, G.G. (1976), Nature 263: 211-214.