Volume 11 Number 24 1983 Nucleic Acids Research

Isolation and characterization of two alleles of the chicken cytochrome *c* gene

Keith J.Limbach and Ray Wu

Section of Biochemistry, Molecular and Cell Biology, Cornell University, Ithaca, NY 14853, USA

Received 9 August 1983; Revised 17 October 1983; Accepted 14 November 1983

ABSTRACT

Analysis of total chicken DNA by genomic blot hybridization indicates that only one cytochrome c gene exists in the chicken genome. The two alleles of this single cytochrome c gene have been isolated from a Charon 4A-chicken genomic library. This isolation made use of the yeast CYC1 cytochrome c gene as a specific hybridization probe. The 2 chicken alleles, CC9 and CC10, have been sequenced. The amino acid sequence predicted by these 2 alleles is identical, and agrees with the published chicken cytochrome c protein sequence. The flanking regions of these 2 alleles exhibit approximately 1% divergence, indicating a very limited polymorphism. Comparative sequence analysis with the flanking regions of previously isolated cytochrome c genes (yeast and rat) indicate no significant regions of homology. The presence of only one cytochrome c-like sequence in the chicken genome is in striking contrast with mammalian genomes, which contain as many as 20-30 cytochrome c-like sequences.

INTRODUCTION

Cytochrome c is a small, heme-containing protein found in the mitochondria of all eucaryotic cells. It is one of a series of proteins involved in the electron transport chain, and is therefore required for aerobic respiration. Due to structural and functional constraints, the cytochrome c proteins of widely divergent organisms exhibit considerable amino acid homology[1]. Assuming this homology to be retained at the DNA level, it was proposed that a cytochrome c gene from one organism could be used as a specific hybridization probe to isolate the cytochrome c genes of other widely divergent organisms[2]. The nucleotide sequences of these cytochrome c genes could then be analyzed from an evolutionary perspective, and useful information regarding the regulation of this gene possibly obtained.

A similar line of investigation has also been pursued for the globin gene family. In fact, a considerable number of globin genes have already been isolated, and certain features, such as introns and regulatory sequences, found to be conserved[3]. Globin genes, however, are not present in all eucaryotic organisms. Consequently, studies involving the evolution of this

gene family are limited. Since the cytochrome c gene is present in all eucaryotic organisms, it would therefore appear to be a better candidate for this type of investigation.

In this communication, we report that the chicken has only one cytochrome c gene. We describe the isolation and characterization of the two alleles of this gene, and compare these sequences with each other and with the previously sequenced cytochrome c genes of yeast and rat. We also examine the relative complexity of cytochrome c-like sequences in chicken and other widely divergent organisms.

MATERIALS AND METHODS

Construction of Chicken Genomic Library

The chicken genomic library, provided by the laboratory of Richard Axel (Columbia University), was constructed in the following manner: Genomic DNA, isolated from the blood of a single Rhode Island Red chicken, was partially digested with the restriction endonucleases HaeIII and AluI, and fragments approximately 15-18 kb long isolated after centrifugation through a sucrose gradient. Synthetic EcoRI linkers were ligated onto the ends of these fragments, which were then ligated onto the purified EcoRI arms of the lambda phage, Charon 4A[4].

Screening of Chicken Genomic Library

From the expected size of the chicken DNA inserts (15-18 kb), and the approximate size of the chicken genome (1.3×10^6 kb), it was estimated that screening approximately 400,000 phage would ensure, with 99% probability, that the entire chicken genome was represented[5]. Phage were propagated in the E. coli strain DP50supF. Techniques involved in the plating and purification of phage, and in the isolation of phage DNA, were performed as described in a detailed protocol from the laboratory of Blattner (University of Wisconsin). Phage DNA was transferred from plaques, denatured and bound to nitrocellulose filters (Schleicher and Schuell) according to the procedure of Benton and Davis[6]. The hybridization probe was the 0.6 kb EcoRI-HindIII fragment of pYC11, which contains the entire coding region of the yeast CYC1 gene, as well as an additional 278 nucleotides of 3' noncoding sequence. This gene was originally isolated[7] using a synthetic oligonucleotide homologous to a portion of the yeast cytochrome c mRNA[8] whose sequence had been deduced from extensive genetic analysis[9]. The probe was labeled by the method of Rigby et al[10], using ^{32}P-labeled dNTP (410 Ci/nmole) from Amersham, DNase I from Sigma, and E. coli DNA polymerase I from Boehringer-Mannheim.

Hybridizations were conducted at 23°C for 30-48 hours in 50% formamide, 5X SSC (0.75 M NaCl, 0.075 M Na_3-citrate), 0.2% SDS, 1X Denhardt's solution[11], 100 µg/ml sonicated calf thymus DNA and ^{32}P-labeled probe (1-4 x 10^5 cpm/filter). Filters were washed at 23°C for 30 min in 50% formamide, 5X SSC, 0.2% SDS (2 times), and 2X SSC, 0.2% SDS (4 times). These hybridization conditions are equivalent to 58°C in 5X SSC in the absence of formamide[12], and favor the stable formation of a 14-18 nucleotide heteroduplex, with an expected T_m of 53-63°C. This calculation is based on the formula $T_m = 81.5 + 16.5 (\log M) + 0.41 (\%G+C) - (650/n)$[13], where n is the heteroduplex length, M is the ionic strength of the hybridization solution (in 5X SSC, M = 0.975 M Na^+) and (% G+C) is 43% (the G+C content of the yeast CYC1 coding region).

Restriction Enzyme Analysis

Restriction endonucleases were purchased from either New England Biolabs or Bethesda Research Laboratories. DNA samples were fractionated by electrophoresis on 0.7-1.0% agarose gels. Restriction fragments were transferred to nitrocellulose filters by the method of Southern[14], and hybridized to ^{32}P-labeled probe as described above.

DNA Sequence Analysis

DNA sequencing was performed by the methods of Maxam and Gilbert[15], and Guo and Wu[16].

DNA and RNA Preparation

Plasmid DNA was isolated according to the alkaline extraction procedure of Birnboim and Doly[17]. Genomic DNA and polyadenylated RNA were isolated in the following manner. The liver of a freshly killed Babcock 300 White Leghorn chicken was immersed in liquid nitrogen, powdered with a mortar and pestle, and homogenized with a Brinkman polytron homogenizer in a 1:1 mixture of 50 mM Tris-HCl (pH 8.5), 10 mM EDTA, 2% SDS and phenol-chloroform-isoamyl alcohol (50:50:1) in a volume of 10 ml per gram of tissue. This suspension was centrifuged at 16,000 x g for 20 minutes. The aqueous layer was extracted 2 more times with phenol-chloroform-isoamyl alcohol (50:50:1), and adjusted to a concentration of 0.1 M sodium acetate (pH 6.0). Nucleic acids were precipitated with 2.5 volumes of ethanol and collected by centrifugation at 7,000 x g for 20 minutes. This precipitate was resuspended in 0.2 M sodium acetate, reprecipitated, and resuspended in water.

Polyadenylated RNA was purified from this nucleic acid mixture by chromatography on oligo(dT)-cellulose (Collaborative Research), as described by Aviv and Leder[18]. DNA was obtained from this mixture after digestion with RNase A. Commercially isolated chicken DNA was purchased from P-L

Biochemicals, Inc.

Hybridization Analysis of Polyadenylated RNA and Genomic DNA

Polyadenylated RNA samples were resolved by electrophoresis on a 1% agarose-formaldehyde gel (electrode buffer: 2.2 M formaldehyde, 0.018 M Na_2HPO_4, 0.002 M NaH_2PO_4). Samples were dissolved in 50% formamide, 2.2 M formaldehyde, 10 mM sodium phosphate (pH 7.0), and heated for 15 min at 50°C before loading. Gel lanes containing size standards were removed, and the bands visualized by staining with ethidium bromide. RNA was transferred to nitrocellulose as described by Thomas[19], and hybridized to ^{32}P-labeled probe as described above.

Hybridizations involving genomic DNA were performed as previously described, however these hybridizations were conducted at 37°C instead of 23°C.

RESULTS

Isolation of Three Putative Chicken Cytochrome c Genes

Yeast and chicken diverged approximately 900 million years (MYs) ago[20]. Due to structural and functional constraints, however, the cytochrome c proteins of these 2 organisms still exhibit considerable amino acid homology. In fact, when the yeast iso-1 and chicken cytochrome c proteins are aligned properly, 5 distinct regions of extensive amino acid homology are observed (Fig. 1). If maximum homology were conserved at the nucleic acid level, these 5 regions would result in homologous blocks of 27, 18, 15, 18 and 33 consecutive base pairs. Due to the degeneracy of the genetic code, it is not possible to predict the actual length of the longest perfect heteroduplex that exists between these 2 genes. However, it was considered likely that a perfect heteroduplex of 14-18 nucleotides would exist. Under the proper conditions, this would represent sufficient nucleotide homology for the interspecies hybridization of the chicken and yeast cytochrome c genes. Based upon this prediction, a Charon 4A-chicken genomic library was screened, using the yeast CYC1 cytochrome c gene as a specific hybridization probe. Since a perfect heteroduplex of 14-18 nucleotides corresponds to a T_m of 53-63°C in 5X SSC[13], all hybridizations were conducted at 23°C in 5X SSC and 50% formamide; conditions which are equivalent to 58°C in 5X SSC.

Approximately 1,200,000 plaques were screened and 38 recombinant phage isolated. The probe for this screening was the 0.6 kb EcoRI-HindIII fragment of pYC11. This fragment can be subdivided by the restriction enzyme TaqI, into a 343 bp fragment containing the entire coding region of the yeast CYC1

Fig. 1. Amino acid sequence of yeast iso-1 and chicken cytochrome c proteins. The yeast iso-1 and chicken cytochrome c amino acid sequences are aligned for maximum homology. Conserved residues are underlined, and 5 regions of extensive amino acid homology are labeled to indicate the relative lengths of the invariant sequence. The amino acid letter code is as follows: A, Ala; R, Arg; N, Asn; D, Asp; C, Cys; E, Glu; G, Gly; H, His; I, Iso; L, Leu; K, Lys; M, Met; F, Phe; P, Pro; S, Ser; T, Thr; W, Trp; Y, Tyr; V, Val.

gene, and a 238 bp fragment containing the 3' noncoding region of this gene.

When the 38 hybridizing recombinant phage were analyzed with the 2 smaller probes, 9 hybridized to the coding region probe and 29 hybridized to the A-T rich 3' noncoding region probe. The coding region isolates were further subdivided into 7 classes by restriction enzyme and hybridization analyses, which revealed that 3 of these clones were identical. The hybridizing fragment of each of these 7 clones was then subcloned into either pBR322 or pBR325, and designated pCC1-pCC7. After further localizing the hybridizing region, preliminary sequencing analysis was performed on 3 of these clones. This analysis revealed that the hybridization of 2 of these clones was the result of a single, fortuitous, 18 bp region of homology to the yeast CYC1 probe. The hybridizing region of the third clone, however, was found to be colinear with a region of the chicken cytochrome c amino acid sequence. This clone (pCC1) was then used as a hybridization probe to examine the other 6 chicken clones (pCC2-pCC7). This analysis revealed that none of the other chicken clones were homologous to the putative chicken cytochrome c gene. Based on these results, no further work was performed on pCC2-pCC7.

The putative chicken cytochrome c gene (CC1) was then used as a hybridization probe to rescreen the chicken library. Approximately 800,000 plaques were screened and 9 recombinant phage isolated. Restriction enzyme and hybridization analyses revealed that these 9 clones could be divided into 3 separate classes (Fig. 2). Two of these classes appeared to be unique, and the hybridizing fragment of each was subcloned into pBR322 and designated pCC9 and pCC10. The third class, however, appeared to be identical to the Charon 4A-CC1 clone.

Of the 9 clones isolated in the second screening, 2 appeared to be identical to the Ch4A-CC1 clone, 2 appeared to be identical to the Ch4A-CC9 clone, and 5 appeared to be identical to the Ch4A-CC10 clone.

Detailed restriction enzyme maps were constructed of pCC1, pCC9 and pCC10 (Fig. 3). The three maps constructed were found to be virtually identical.

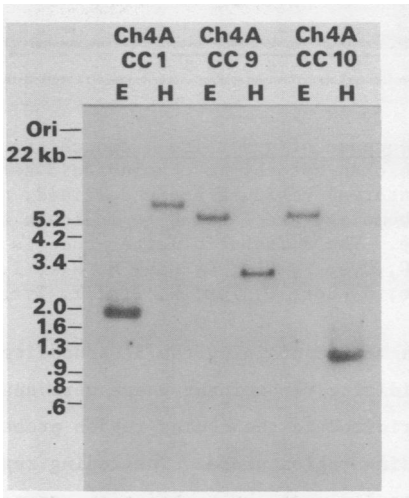


Fig. 2. Hybridization analysis of Ch4A-CC1, Ch4A-CC9 and Ch4A-CC10 DNA. Each lane contains 1 µg of DNA isolated from Ch4A-CC1, Ch4A-CC9 or Ch4A-CC10. Samples were digested with EcoRI (E) or HindIII (H), transferred to nitrocellulose and hybridized to the ^{32}P-labeled EcoRI-HindIII fragment of pYC11 (1 x 10^7 cpm/µg). Size standards are the EcoRI-HindIII digestion fragments of λ.

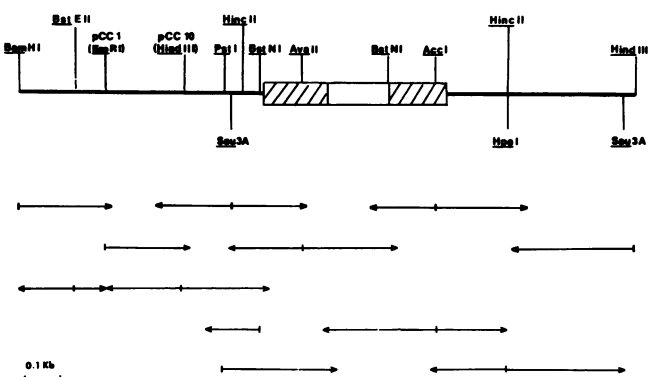


Fig. 3. Partial restriction enzyme map and DNA sequencing strategy of pCC1, pCC9 and pCC10. The direction of transcription is from left to right. Exons and introns are represented as crosshatched and open boxes, respectively. Brackets around the EcoRI site indicate that this site is unique for pCC1. Brackets around the HindIII site indicate that this site is unique for pCC10. The arrows below indicate the nucleotide sequencing strategy used. pCC9, pCC10 and certain regions of pCC1 were sequenced by the method of Maxam and Gilbert, as represented by the arrows in the top 3 lines. The majority of pCC1 was sequenced by the method of Guo and Wu, as represented by the arrows in the bottom 2 lines.

In fact, if pCC9 is taken as a standard, the only difference in the 3 maps is the additional EcoRI site in pCC1, and the additional HindIII site in pCC10.

Sequence Analysis of the Three Putative Chicken Cytochrome c Genes

Nucleotide Sequence Differences. Sequence analysis of pCC1, pCC9 and pCC10 were performed by the method of Maxam and Gilbert[15], and Guo and Wu[16] according to the strategy depicted in Fig. 3. This analysis revealed that the nucleotide sequence of pCC1 and pCC9 were identical. Further restriction enzyme analysis of Charon 4A-CC1 and Charon 4A-CC9 revealed that these 2 recombinant phage contain overlapping fragments of the same chicken cytochrome c gene. Restriction enzyme analysis of Ch4A-CC1, for instance, showed that the hybridizing region of this clone was at the extreme edge of the cloned chicken DNA. The additional EcoRI site found in this clone was therefore, not due to chicken DNA, but was created during the construction of the genomic library, when EcoRI linkers were ligated onto the ends of partially digested chicken DNA.

This sequence analysis also revealed that the nucleotide sequence of pCC9 (pCC1) and pCC10 were highly homologous (Fig. 4). Over 1600 bp were sequenced in both clones and only 20 nucleotide differences were observed. Eleven of these differences occur in the 5' noncoding sequence, two occur in the 3' noncoding sequence and seven occur in an apparent intervening sequence. Since none of these differences occur within the coding region, the cytochrome c proteins encoded by these two genes should be identical.

Twelve of these differences are the result of single-base-pair substitution events and 8 are the result of 2 individual insertion (deletion) events. Both of the insertion (deletion) events occur in a relatively long stretch of T's. Relative to the CC9 gene, a 2 bp deletion and a 6 bp insertion have occurred in CC10. The deletion occurs at position -265, and the insertion occurs at position +317.

Of the 12 single-base-pair substitution events, 8 are transitions and 4 are transversions. This is consistent with data presented by Gojobori et al[21], which indicate that at selectively neutral positions, transition mutations occur more frequently than transversion mutations. Gojobori et al also concluded that the most frequent substitution is a C → T change, and that, in general, substitutions are more likely to occur at a C or G than at a T or A.

Analysis of the 2 chicken genes revealed that 8 of the 12 substitutions from CC10 → CC9 result in a C or G → T or A change, and that 4 of these substitutions result in a C → T change. In contrast, only 4 of the 12

```
CC10   A
CC9  CAGCGGAAGG TCACGGCCGG GCCTAGCGCC TCAGGGGTGC CGCGTCCCGA TGGGCCCGGG GGGGTGCGGG AGCCTCCCTG
     C
     TGGACAAAGC TGCCGGGCCG GCCCTCGTCC CCTCGTGAGG TGACCGCGCC GGGCGCCCGC TCGCCTCAAG GTCAGCCTCC

     GCGTGCCCCG CCTGGGCAGC AGCTGGGCCT TGCGGCACTG CCCTCGGCGC TCCGAGGCGT GTGCTGCCCT GCGGGGCGGA
                                                                  A
     ACGTGAGCAG TTCTGCCGTT TACGCGAGGC TGTTGGTGAC TGGGATGCCC CGCGGTTATT GATGCCAAAG CCGTACGTAG

     AACTTGCTCT CAATTTGCAC TTCAGAAGTT ATTGCCGTGC TTCTTAAATG CTCTGGGGGA ATTACTTGAG CTGAGGCATG
        C(--)                                  C          G          G     C
     TCTTTTTTTT TTTTTCCCTG TTCTTCGGGC ATAAAGTTTC ACCACATAAT CTTTTACATA GTTTAATGGT TATATGCAAA
                                                                  A          -200
     TCTTCAAACC CAAGAGCTAT AACTCGTGGC TGGAAGGCAT TCAGGAGAGT TACCAAATTG CTGGACTGTA GCTGCAGTTT

     TGTAAACCAG TGTGGATCTT TTTAAAGTTT CTGTTGAGTA GTTGACCTTG ATATTGTGAT ACTGACCTCT GTGTGTGTTC
    -100
                          1
                      Met Gly Asp Ile Glu Lys Gly Lys Lys Ile Phe Val Gln Lys Cys Ser Gln
     CTGGTCTAGT ACTGACA ATG GGA GAT ATT GAG AAG GGC AAG AAG ATT TTT GTC CAG AAA TGT TCC CAG
                        -1 1
     Cys His Thr Val Glu Lys Gly Gly Lys His Lys Thr Gly Pro Asn Leu His Gly Leu Phe Gly Arg
     TGC CAT ACG GTT GAA AAA GGA GGC AAG CAC AAG ACT GGA CCC AAC CTT CAT GGC CTG TTT GGA CGC
                                                                  55
     Lys Thr Gly Gln Ala Glu Gly Phe Ser Tyr Thr Asp Ala Asn Lys Asn Lys
     AAA ACA GGA CAA GCT GAG GGC TTC TCT TAC ACA GAT GCC AAT AAG AAC AAA G GTAAATGTAA
                                                                     166

     AGCAGCTTTC CTGAGCCACT GGTGGTTCTT GGTGATAGTC AGAAGTACTG AAACCTGTAG AAAGCAAAAT AATATGAATG

            A                                                          (TTTTTA)
     ACCAGAAACA AATGTATTCA TTTCACTTTG AAGTCACCGC TTGTTTTTTT TCCTGTTTTT T------TTT ATTTTCTTGC

        Gly Ile Thr Trp Gly Glu Asp Thr Leu Met Glu Tyr Leu Glu Asn Pro Lys Lys Tyr Ile Pro
     CAG GT ATC ACT TGG GGT GAG GAT ACT CTG ATG GAG TAT TTG GAA AAT CCA AAG AAG TAC ATC CCA

     Gly Thr Lys Met Ile Phe Ala Gly Ile Lys Lys Lys Ser Glu Arg Val Asp Leu Ile Ala Tyr Leu
     GGA ACA AAG ATG ATT TTT GCG GGT ATC AAG AAG AAG TCT GAG AGA GTA GAC TTA ATA GCA TAT CTC
                          104
     Lys Asp Ala Thr Ser Lys
     AAA GAT GCC ACT TCA AAG TAA AAGTTAT CTGCTGCCTT ATTTATTTCA CAAAGGAGAT GGCAATGGAA
                              482                 500
     GTGTCTGTGA CAAGATTGGT TTTTAAACTT TCTATTTTTA CATATACCAT GTCTAACCTT AAAATCGGTT TTACCCATCA
                                                                                  600
     GATAATGTTG CTCATGATGG GTCACTGAAG AACACACTTG GCAGCCGTTA ACTTACGGAA ACTATGTAAC TGGGTTGATT
           G                                                      G
     TAAATGAATA TAATGTTCAG TTGTTCTTAA TCATGAAATT AAAAACAAAA AAAAATTTCC TGCCTCTTCA TTGTTTAAAA
               700
     CAAAAAAAGT GTCAAACAAG TGTCAAAAGT GTAATAGCTT TGACAACCCA ACTCTTCAAA ATAAAAAAGT AGAGTGGGTC
                                      800
     TTGGCATCTT CCATCATAAG TCTTCTGTTA CATTACACTA AAGATTTAAA AGTAAACAAA GACTTAAAAA ATATTTTTCT

     AAACTTGATC AGTTCTGGGT GTTGTGG
```

substitutions from CC9 → CC10 result in a C or G → T or A change, and none result in a C → T change. Assuming the conclusions of Gojobori et al to be correct, the above data would indicate that the CC10 gene bears more resemblance to the ancestral cytochrome c sequence than the CC9 gene.

One of the twelve substitution events occurs at position -224, and results in the creation of an additional HindIII site in pCC10. This unique HindIII site changes the size of the HindIII hybridizing fragment of this clone, and therefore allows these 2 nearly identical genes (CC9 and CC10) to be easily distinguished.

In a study of insulin and globin genes, Perler et al[22] estimated the rate of silent site mutations to be approximately 7×10^{-9} substitutions per nucleotide per year. Using this estimate, the 2 chicken genes were calculated to have diverged approximately 1.5 MYs ago. (14 substitutions/1300 nucleotides/7×10^{-9} substitutions/nucleotide/year = 1,500,000 years.) This calculation utilizes a value of 1300 to represent the number of nucleotide sites at which neutral substitutions could possibly occur. This value represents the entire nucleotide sequence shown in Fig. 4, excluding the coding region. Since Perler et al count each deletion or insertion as a single mutational event, a value of 14 is used to represent the number of substitution events which have occurred between these 2 genes.

Coding, Noncoding and Intervening Sequences. The amino acid sequence predicted by CC9 and CC10 agrees exactly with the chicken cytochrome c protein sequence, as determined by Chan and Margoliash[23]. Colinearity with the amino acid sequence is interrupted once, however, at residue 56, by an apparent intervening sequence. The intron-exon junctions of this putative intervening sequence follow the obligatory GT/AG rule, and in general, agree quite well with the consensus RNA splice site sequences[24]. The model RNA splice donor sequence ($AG^{\downarrow}GT^{G}_{A}AGT$), for instance, agrees quite well with the chicken sequence ($AG^{\downarrow}GTAAAT$) found at the point where colinearity between the nucleotide and amino acid sequences diverge. Additionally, the consensus RNA splice acceptor sequence ($PyNPyPyPyNCAG^{\downarrow}GT$) agrees well with the chicken sequence ($TCTTGCCAG^{\downarrow}GT$) found at the point where colinearity with the amino

Fig. 4. Nucleotide sequences of two chicken cytochrome c genes. The nucleotide sequences of the CC9 and CC10 genes are presented. Only those nucleotides that differ from CC9 are shown for CC10. The amino acid sequence predicted by these genes is displayed on the line above the nucleotide sequence. The numbering system begins from the initiation codon, with nucleotides being numbered below the DNA sequence, and amino acid residues being numbered above. Putative control sequences are underlined.

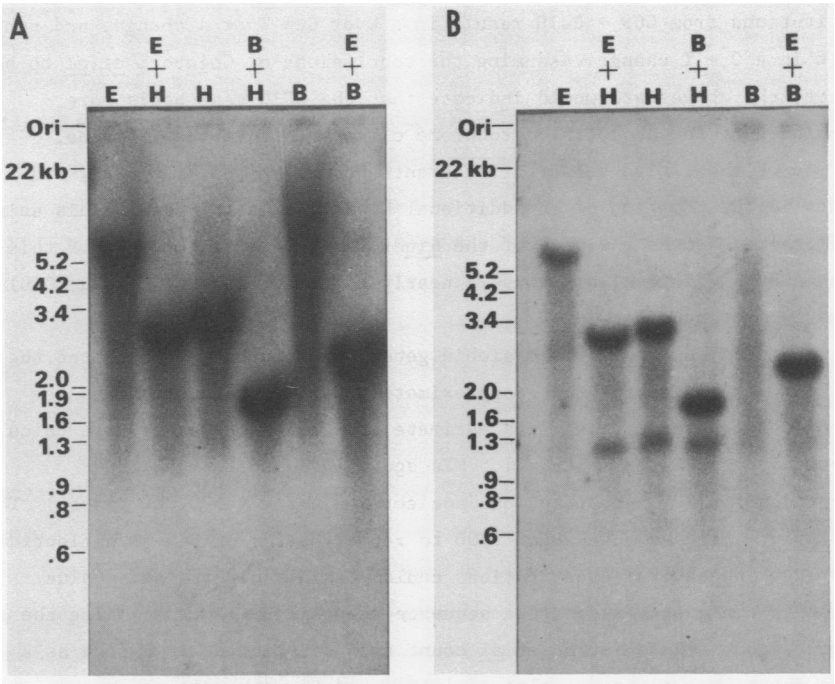


Fig. 5. Hybridization analysis of chicken genomic DNA. (A) Each lane contains 40 µg of genomic DNA isolated from a single White Leghorn chicken. (B) Each lane contains 40 µg of commercially isolated chicken genomic DNA. Samples were digested with EcoRI (E), HindIII (H) and/or BamHI (B), transferred to nitrocellulose and hybridized to the ^{32}P- labeled PstI-HindIII fragment of pCC9 (5 x 10^7 cpm/µg). In Fig. 5A, the autoradiogram was purposely overexposed in order to increase the chance of detecting any additional hybridizing bands. Size standards were the same as in Fig. 2.

acid sequence resumes. From this information it can be concluded that CC9 and CC10 are interrupted at the first nucleotide of codon 56, by intervening sequences of 167 bp and 173 bp, respectively.

Control Regions. In addition to an intervening sequence, these two chicken genes also have 2 other features common to eucaryotic genes; the promoter sequence, TATAA, and the hexanucleotide polyadenylation sequence, AATAAA.

The promoter sequence, TATAA, is usually found approximately 30 bp upstream from the start of mRNA transcription[25]. This sequence is found in the 2 chicken genes at position -160. No ATG codons occur between this putative promoter sequence and the correct initiation codon. Therefore, assuming this to be the proper promoter sequence, translation of the 2 chicken genes would

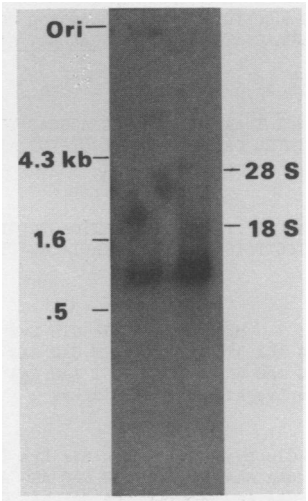
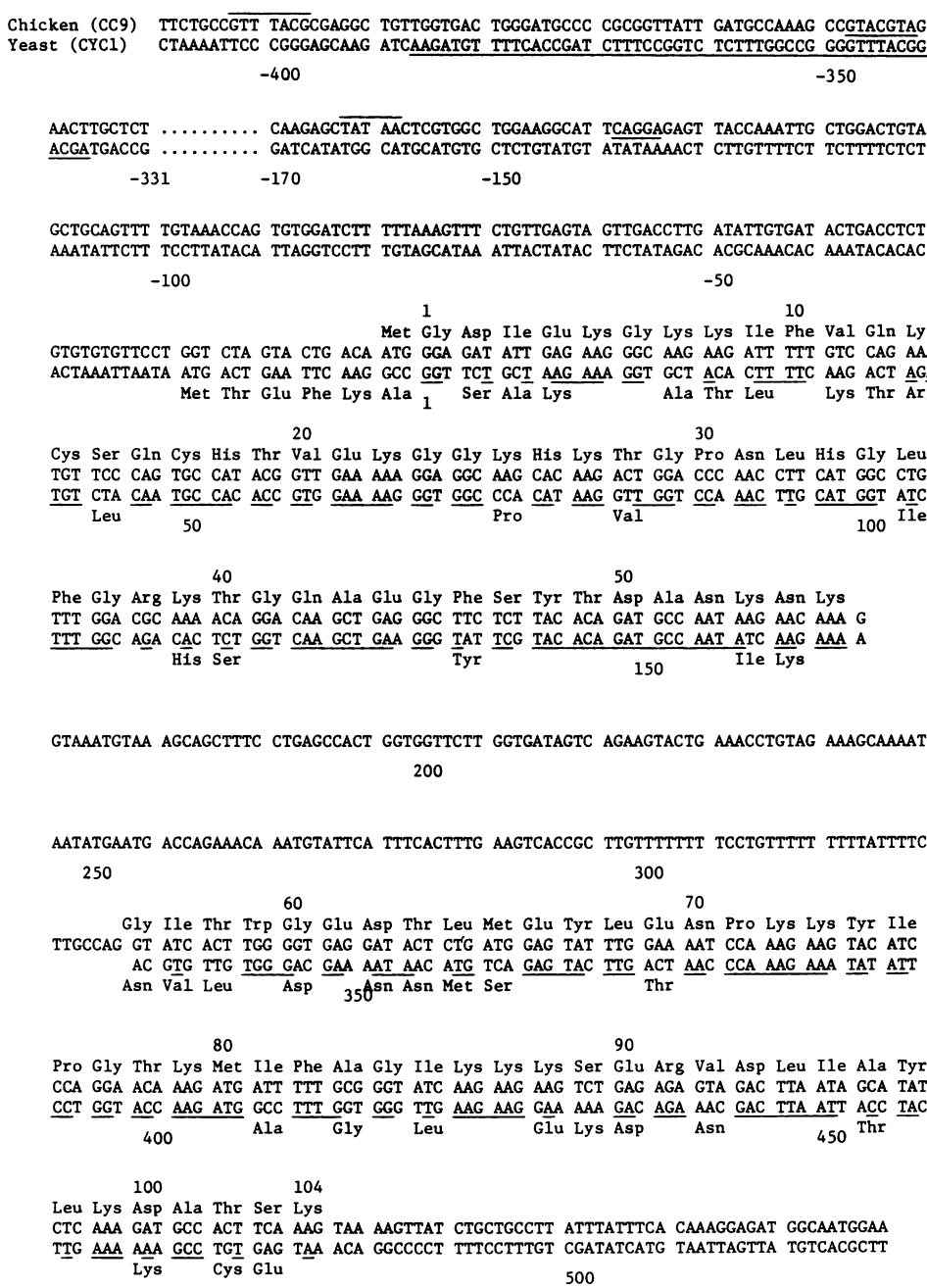


Fig. 6. Hybridization analysis of chicken polyadenylated RNA. 5 µg and 15 µg of polyadenylated RNA was resolved by gel electrophoresis, transferred to nitrocellulose and hybridized to the ^{32}P-labeled PstI-HindIII fragment of pCC9 (5 x 10^7 cpm/µg). Size standards include 28S and 18S chicken ribosomal RNA, EcoRI digested pBR322 and HinfI digested pBR322.

begin at the AUG codon closest to the 5' end of the mRNA molecule. This would be consistent with data by Kozak[26], which indicates that translation initiation usually occurs at the first AUG codon of the mRNA molecule.

The hexanucleotide sequence, AATAAA, is commonly found approximately 20 bp upstream from the polyA addition site[27]. This sequence is found in the 2 chicken genes at position +830, approximately 350 bp downstream from the TAA termination codon.

Base Composition of 5' and 3' Noncoding Sequences. The G + C content of the 3' noncoding sequence of the 2 chicken genes is approximately 32%. This is not unusually low because the coding region of many eucaryotic genes is followed by a region of low G + C content. By contrast, regions of high G + C content are relatively rare. Nevertheless, a region of unusually high G + C content exists in the 5' noncoding sequence of both chicken cytochrome c genes. This region stretches from position -860 to -360 (not all data shown), and has a G + C content of 74%. The significance of this G-C rich region is not yet known, however, it does explain why at least 3 of the 10 cytochrome c-like sequences isolated from this Charon 4A-chicken genomic library contain chicken DNA inserts which begin in this region. These 3 clones, which represent the Ch4A-CC1 class of recombinant phage isolates, contain an EcoRI


Chicken (CC9) TTCTGCCGTT TACGCGAGGC TGTTGGTGAC TGGGATGCCC CGCGGTTATT GATGCCAAAG CCGTACGTAG
Yeast (CYC1) CTAAAATTCC CGGGAGCAAG ATCAAGATGT TTTCACCGAT CTTTCCGGTC TCTTTGGCCG GGGTTTACGG
-400
-350
AACTTGCTCT ..... CAAGAGCTAT AACTCGTGGC TGGAAGGCAT TCAGGAGAGT TACCAAATTG CTGGACTGTA
ACGATGACCG ..... GATCATATGG CATGCATGTG CTCTGTATGT ATATAAAACT CTTGTTTTCT TCTTTTCTCT
-331
-170
-150
GCTGCAGTTT TGTAAACCAG TGTGGATCTT TTTAAAGTTT CTGTTGAGTA GTTGACCTTG ATATTGTGAT ACTGACCTCT
AAATATTCTT TCCTTATACA TTAGGTCCTT TGTAGCATAA ATTACTATAC TTCTATAGAC ACGCAAACAC AAATACACAC
-100
-50
1
10
Met Gly Asp Ile Glu Lys Gly Lys Lys Ile Phe Val Gln Lys
GTGTGTGTTCCT GGT CTA GTA CTG ACA ATG GGA GAT ATT GAG AAG GGC AAG AAG ATT TTT GTC CAG AAA
ACTAAATTAATA ATG ACT GAA TTC AAG GCC GGT TCT GCT AAG AAA GGT GCT ACA CTT TTC AAG ACT AGA
Met Thr Glu Phe Lys Ala 1 Ser Ala Lys Ala Thr Leu Lys Thr Arg
20
30
Cys Ser Gln Cys His Thr Val Glu Lys Gly Gly Lys His Lys Thr Gly Pro Asn Leu His Gly Leu
TGT TCC CAG TGC CAT ACG GTT GAA AAA GGA GGC AAG CAC AAG ACT GGA CCC AAC CTT CAT GGC CTG
TGT CTA CAA TGC CAC ACC GTG GAA AAG GGT GGC CCA CAT AAG GTT GGT CCA AAC TTG CAT GGT ATC
Leu 50 Pro Val 100 Ile
40
50
Phe Gly Arg Lys Thr Gly Gln Ala Glu Gly Phe Ser Tyr Thr Asp Ala Asn Lys Asn Lys
TTT GGA CGC AAA ACA GGA CAA GCT GAG GGC TTC TCT TAC ACA GAT GCC AAT AAG AAC AAA G
TTT GGC AGA CAC TCT GGT CAA GCT GAA GGG TAT TCG TAC ACA GAT GCC AAT ATC AAG AAA A
His Ser Tyr Ile Lys
150
GTAAATGTAA AGCAGCTTTC CTGAGCCACT GGTGGTTCTT GGTGATAGTC AGAAGTACTG AAACCTGTAG AAAGCAAAAT
200
AATATGAATG ACCAGAAACA AATGTATTCA TTTCACTTTG AAGTCACCGC TTGTTTTTTT TCCTGTTTTT TTTTATTTTC
250
300
60
70
Gly Ile Thr Trp Gly Glu Asp Thr Leu Met Glu Tyr Leu Glu Asn Pro Lys Lys Tyr Ile
TTGCCAG GT ATC ACT TGG GGT GAG GAT ACT CTG ATG GAG TAT TTG GAA AAT CCA AAG AAG TAC ATC
AC GTG TTG TGG GAC GAA AAT AAC ATG TCA GAG TAC TTG ACT AAC CCA AAG AAA TAT ATT
Asn Val Leu Asp 350 Asn Asn Met Ser Thr
80
90
Pro Gly Thr Lys Met Ile Phe Ala Gly Ile Lys Lys Lys Ser Glu Arg Val Asp Leu Ile Ala Tyr
CCA GGA ACA AAG ATG ATT TTT GCG GGT ATC AAG AAG AAG TCT GAG AGA GTA GAC TTA ATA GCA TAT
CCT GGT ACC AAG ATG GCC TTT GGT GGG TTG AAG AAG GAA AAA GAC AGA AAC GAC TTA ATT ACC TAC
400 Ala Gly Leu Glu Lys Asp Asn 450 Thr
100
104
Leu Lys Asp Ala Thr Ser Lys
CTC AAA GAT GCC ACT TCA AAG TAA AAGTTAT CTGCTGCCTT ATTTATTTCA CAAAGGAGAT GGCAATGGAA
TTG AAA AAA GCC TGT GAG TAA ACA GGCCCCT TTTCCTTTGT CGATATCATG TAATTAGTTA TGTCACGCTT
Lys Cys Glu
500

linker at virtually identical positions because this relatively small, 500 bp G-C rich region of DNA contains an unusually high number (11) of HaeIII sites. Since this library was constructed by partially digesting chicken DNA with HaeIII and AluI, a much greater than random percentage of chicken DNA inserts would therefore be expected to begin in this region.

Complexity of Cytochrome c-like Sequences in the Chicken Genome

In order to investigate the complexity of cytochrome c-like sequences in the chicken genome, total DNA was isolated from an individual chicken, cut with various restriction endonucleases, fractionated by electrophoresis, transferred to nitrocellulose and hybridized to the PstI-HindIII fragment of pCC9 (Fig. 5A). Only one hybridizing fragment was observed in each lane, indicating that only one cytochrome c-like sequence was present in the genome of this particular chicken. The size of these hybridizing fragments correspond exactly with the size of the restriction enzyme products expected from digestion of the Ch4A-CC9 chicken gene (Fig. 2 and Fig. 3). This analysis was repeated using genomic DNA isolated from another individual chicken and the results were identical. This analysis was then repeated using commercially isolated chicken DNA. In this experiment, 2 hybridizing fragments were observed in those lanes in which the DNA was cut with HindIII (Fig. 5B). The size of these 2 hybridizing fragments correspond exactly with the size of the restriction enzyme products expected from digestion of the Ch4A-CC9 and Ch4A-CC10 genes (Fig. 2 and Fig. 3). These 2 hybridizing fragments, however, were only observed in those lanes in which the DNA was cut with HindIII. In those lanes in which the DNA was cut with enzymes other than HindIII, only 1 hybridizing fragment was observed. This result is consistent with restriction enzyme and sequencing data, which indicated that except for the additional HindIII site in CC10, the 2 chicken cytochrome c genes were extremely similar. In conclusion, it appears that both chicken cytochrome c

Fig. 7. Nucleotide sequences of chicken and yeast CYC1 cytochrome c genes. The nucleotide sequences of the chicken and yeast CYC1 cytochrome c genes are aligned for maximum homology. The predicted amino acid sequence of the chicken gene is displayed on the line above the nucleotide sequences. Any amino acid differences in the yeast gene are indicated on the line below these sequences. Nucleotide homology within the coding regions of these 2 genes is indicated by a horizontal line underneath the 2 sequences. Putative promoter sequences for the yeast and chicken genes are overlined. The putative heme regulatory region of the yeast gene is underlined (-387 to -337). Homology to the chicken sequence within this heme regulatory region is overlined (-403 to -397 in chicken, -348 to -342 in yeast). The nucleotide sequence of the yeast CYC1 gene (-410 to -331) was provided by Michael Smith (personal communication).

genes (CC9 and CC10) are present in the commercially isolated DNA, but only the CC9 gene is present in the DNA of the 2 individual chickens analyzed.

The relative intensities of the 2 hybridizing bands, in the commercially isolated chicken DNA, are not equivalent. The larger band (CC9) is approximately 10 times more intense than the smaller band (CC10). Therefore, in the commercial DNA sample analyzed, the CC9 gene is approximately 10 times more abundant than the CC10 gene. It should be emphasized that the source of this commercially isolated DNA was not a single individual chicken, but rather, a heterogeneous population of chickens. The fact that 2 individual chickens contain only one of the 2 cytochrome c genes, whereas a population of chickens contain both of these genes, suggests that CC9 and CC10 are not 2 individual cytochrome c genes, but rather, allelic forms of the same cytochrome c gene.

Assuming this to be true, we can infer that the 2 individual chickens from which genomic DNA was isolated contained 2 copies of the CC9 allele. We can also conclude that the particular chicken used to construct the Charon 4A-chicken genomic library contained both the CC9 and CC10 alleles. The latter conclusion is supported by the relative frequency of cytochrome c alleles isolated from this library. Overall, 10 clones were isolated; 3 were like Ch4A-CC1, 2 like Ch4A-CC9 and 5 like Ch4A-CC10. Since Ch4A-CC1 and Ch4A-CC9 contain the same allele, both cytochrome c alleles were isolated 5 times. Therefore, in this particular library, the relative frequency of these 2 alleles appears to be equal.

Analysis of Chicken Cytochrome c mRNA

To estimate the number and size of chicken cytochrome c mRNAs, polyadenylated RNA was isolated from an individual chicken, fractionated on an agarose-formaldehyde gel, transferred to nitrocellulose and hybridized to the PstI-HindIII fragment of pCC9 (Fig. 6). A single hybridizing species, approximately 900 to 950 nucleotides long, was observed. The size of this observed mRNA species agrees quite well with the size of the mRNA molecule estimated from the DNA sequence. This estimated mRNA molecule assumes that transcription starts approximately 30 nucleotides from the putative promoter sequence (at -160), and ends approximately 20 nucleotides from the putative polyadenylation site (at +830). By subtracting the length of the intron and allowing an additional 110 nucleotides for the polyA tail, the resultant mRNA molecule would be approximately 920 nucleotides long, which is in good agreement with the observed value.

DISCUSSION

Complexity of Cytochrome c-like Sequences in Chicken, Yeast and Mammalian Genomes

We have shown that the chicken contains only one functional cytochrome c gene per haploid genome. The yeast Saccharomyces cerevisiae contains 2 functional genes per haploid genome[2], CYC1[28] and CYC7[2]. In both cases, these functional genes represent the only cytochrome c-like sequences found in the genomes of these 2 species. In striking contrast, mammals seem to contain as many as 20-30 cytochrome c-like sequences, many of which appear to be non-functional. Hybridization analysis of total rat DNA, for instance, revealed approximately 25 cytochrome c-like genomic fragments[29]. The probe for this analysis was the putative rat cytochrome c gene of pRC4[30]. This gene was also used as a hybridization probe to analyze the genomic DNA of other mammals, such as canine, bovine and human. In each mammalian organism analyzed, approximately 20-30 hybridizing genomic fragments were observed[29]. It therefore appears that a vastly increased number of cytochrome c-like sequences is a general feature of mammalian genomes.

Three of these cytochrome c-like fragments have been sequenced in rat. All 3 appear to be pseudogenes, and all 3 appear to have arisen via an insertion into the genome of a cDNA copy of a cytochrome c mRNA molecule[31]. These results indicate that the increase in cytochrome c-like sequences in mammalian genomes may be due to an increase in pseudogenes, which have arisen via the reverse transcription of mRNA molecules. These results also suggest the possibility that a mechanism to synthesize and insert cDNA copies of mRNA molecules into the genome exists in mammals, but not in chickens or yeast.

Comparison of the Yeast, Rat and Chicken Cytochrome c Genes

Cytochrome c genes from four species have now been isolated and sequenced; the CYC1[28] and CYC7[2] genes of the yeast S. cerevisiae, the PoCYC gene of the fission yeast Schizosaccharomyces pombe[32], the RC4[30] gene of rat, and the CC9 (CC10) alleles of chicken. The remainder of this report will attempt to draw some preliminary conclusions from a comparison of these sequences.

Coding Region. The overall nucleotide sequence homology between the coding regions of the yeast CYC1 and chicken cytochrome c genes is 62%. As expected, much of this homology is concentrated in the regions encoding the 5 blocks of extensive amino acid homology (Fig. 1). The longest stretch of perfect DNA homology between these 2 genes is 16 nucleotides. This homology (nucleotides 142-157) is not found in the longest block of invariant amino

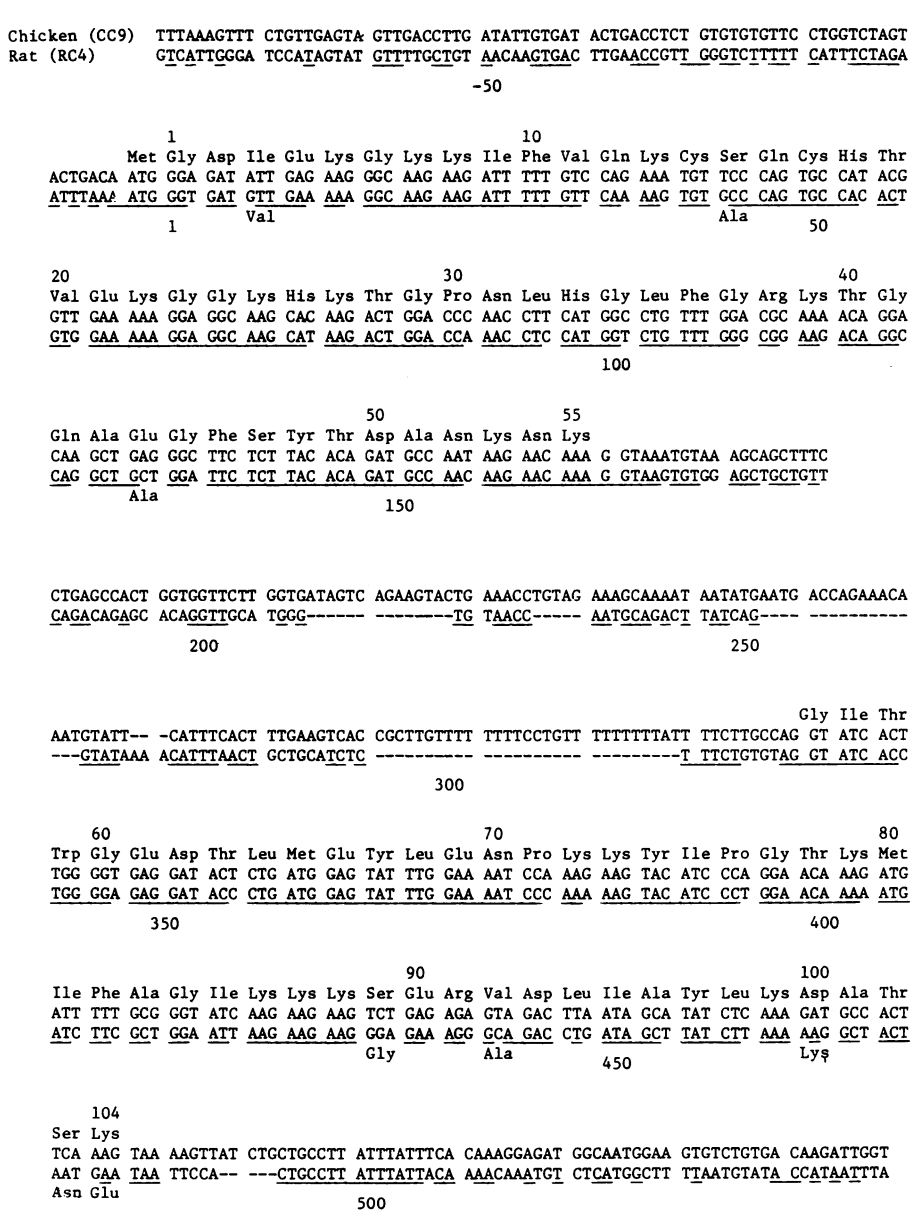


Fig. 8. Nucleotide sequences of chicken and rat cytochrome c genes. The nucleotide sequences of the chicken and the rat genes are aligned for maximum homology. The predicted amino acid sequence of the chicken gene is displayed on the line above the nucleotide sequences. Any amino acid differences in the rat gene are indicated on the line below these sequences. Nucleotide homology between these 2 genes is indicated by a horizontal line underneath the 2 sequences.

acids, but is found in the relatively small block between amino acids 48-53 (Fig. 7).

Regulatory Sequences. It was hoped that the sequences necessary for the proper regulation of the cytochrome c gene would be conserved, and that the evolutionary development of these regions could be examined. However, comparing the 5' and 3' noncoding sequences of the 5 isolated cytochrome c genes, no consistent pattern of conservation is observed. In fact, even some of the putative control regions thought to be ubiquitous for eucaryotic genes are not conserved in all 5 cytochrome c genes. For example, the promoter sequence, TATAA, is present in only 2 of the 5 cytochrome c genes. It is found approximately 160 bp from the start of the chicken gene and about 120 bp from the start of the yeast CYC1 gene (Fig. 7). This sequence, however, is not found in either the rat RC4 gene, the yeast CYC7 gene or the yeast PoCYC gene.

The lack of a promoter sequence has been used to explain the under-expression of the yeast CYC7 gene. It has been hypothesized that the absence of a promoter sequence results in the inefficient transcription initiation of the CYC7 gene, and therefore explains why this gene contributes only 5-10% of the total cytochrome c of the yeast cell. On the other hand, it is not known whether the expression of the rat RC4 gene or the yeast PoCYC gene are affected by the apparent lack of a promoter sequence.

Another regulatory sequence found in the majority of eucaryotic genes is the hexanucleotide polyadenylation sequence, AATAAA. Once again this sequence is found in only 2 of the 4 cytochrome c genes of which the 3' flanking sequence has been determined. However, in this case, the regulatory sequence is found in the chicken and the rat genes, but not in the yeast CYC1 or CYC7 genes. Although only 4 genes have been analyzed, this data may indicate a trend towards the use of this regulatory sequence in the cytochrome c genes of higher eucaryotes. Although the hexanucleotide sequence is conserved in both the chicken and the rat genes, there still appears to be a significant difference between the transcription termination of these 2 genes. Transcription of the chicken gene appears to terminate at one definite site, resulting in one mRNA species. The transcription of the rat RC4 gene, however, appears to terminate at 3 different sites, resulting in 3 different mRNA species[31]. Each of these rat transcripts does appear to terminate approximately 20 nucleotides downstream from an AATAAA sequence, however, since the termination of the chicken and the yeast CYC1[33] cytochrome c genes occur at only one site, this additional regulation of the rat RC4 gene does

not appear to apply to all cytochrome c genes.

Another regulatory element, found in the yeast CYC1 gene, is the heme regulatory site. By fusing the upstream region of the yeast CYC1 gene to the E. coli lacZ gene, and then studying the levels of ß-galactosidase in heme-deficient cells, Guarente and Mason[34] showed that the expression of the yeast CYC1 gene is tightly regulated by levels of intracellular heme. Subsequent deletion analysis identified a heme regulatory site, located within a 50 bp sequence centered approximately 275 bp upstream from the site of transcription initiation. It is not known whether other eucaryotic cytochrome c genes are regulated by intracellular heme levels, however, preliminary experiments with the yeast CYC7 gene seem to indicate that this gene is not under heme control[35].

Since the sequence of the yeast CYC1 heme regulatory region is known, the upstream region of the chicken cytochrome c gene was analyzed for possible homology to this regulatory region. This analysis revealed one small region of homology. Approximately 270 bp from the putative start of chicken cytochrome c transcription, a 7 bp sequence was found which is identical to a portion of the yeast CYC1 heme regulatory region (Fig. 7). Although this chicken sequence is located in the same approximate position as the yeast sequence, this homologous region is relatively short and may simply be fortuitous. Whether the chicken gene is actually regulated by heme levels remains to be determined.

5' and 3' Noncoding Regions. The 5' and 3' noncoding regions of the 5 isolated cytochrome c genes exhibit very little homology. One possible exception, however, occurs in the 3' noncoding region of the chicken, rat and yeast CYC7 genes. Located 10 bp from the chicken termination codon is the sequence, CTGCCTTATTTATT. This 14 bp sequence is also found located 5 bp from the rat termination codon (Fig. 8). In addition, the final 8 bp of this sequence, TATTTATT, is also found near the termination codon of the yeast CYC7 gene. In fact, this latter homology does not need to be shifted to be properly aligned with the chicken sequence. The significance of this apparent homology is not yet known, especially since this sequence is absent from both the yeast CYC1 and PoCYC genes.

Intervening Sequences. Of the 5 cytochrome c genes isolated, only the chicken and the rat genes contain an intervening sequence (Fig. 8). This intron interrupts both genes at the identical position and therefore appears to be relatively stable.

It has been observed that the introns of different genes exhibit

different relative stabilities. For example, many actin genes have been isolated in which the number and position of introns is not conserved. Even in the same organism, different actin genes have been isolated which contain introns at different positions[36]. Conversely, every functional globin gene so far isolated (excluding the plant leghemoglobin gene) contains 2 intervening sequences at homologous positions[3]. Since the α and β globin genes are thought to have diverged approximately 500 MYs ago[20], these 2 intervening sequences must have been present in the ancestral globin sequence before this divergence occurred. The plant leghemoglobin gene contains a third intervening sequence between amino acids 68 and 69[37], however, since the plant-animal divergence occurred several hundred MYs before the α-β globin divergence, it can still be concluded that the globin introns are extremely stable.

The intervening sequence of the chicken CC9 cytochrome c gene is 167 bp long. The intervening sequence of the rat RC4 gene is 105 bp long. The homology between these 2 introns is shown in Fig. 8. Gaps have been added to maximize homology. The large amount of divergence between these 2 introns in consistent with data which estimates that birds and mammals diverged approximately 300 MYs ago.

ACKNOWLEDGEMENTS

We would like to thank Elizabeth Keller, Joseph Calvo and J. Yun Tso for their valuable criticisms of the manuscript, and Fran Ormsbee for typing the manuscript. This work was supported by research grant GM29179 from the National Institutes of Health, U.S. Public Health Service.

REFRENCES

1. Margoliash, E. and Schejter, A. (1966) Adv. Protein Chem. 21, 114-286.
2. Montgomery, D.L., Leung, D.W., Smith, M., Shalt, P., Faye, G. and Hall, B.D. (1980) Proc. Natl. Acad. Sci. USA 77, 541-545.
3. Efstratiadis, A., Posakony, 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., Smithies, O., Baralle, F., Shoulders, C. and Proudfoot, N.J. (1980) Cell 21, 653-668.
4. Blattner, F.R., Williams, B.G., Blechl, A.E., Thompson, K.D., Faber, H.E., Furlong, L.A, Grunwald, D.J., Kieter, D.O., Moore, D.D., Schumm, J.W., Sheldon, E.L. and Smithies, O. (1977) Science 196, 161-169.
5. Clarke, L. and Carbon, J. (1976) Cell 9, 91-99.
6. Benton, W.D. and Davis, R.W. (1977) Science 196, 180-182.
7. Montgomery, D., Hall, B., Gillam, S. and Smith, M. (1978) Cell 14, 673-680.
8. Szostak, J.W., Stiles, J.I., Bahl, C.P. and Wu, R. (1977) Nature 265, 61-63.
9. Stewart, J.W. and Sherman, F. (1974) in Molecular and Environmental

Aspects of Mutagenesis (Prakash, L., Sherman, F., Miller, M., Lawrence, C. and Taber, H., eds.), pp. 102-127, Charles Thomas, Springfield, Ill.
10. Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113, 237-251.
11. Denhardt, D.T. (1966) Biochem. Biophys. Res. Commun. 23, 641-646.
12. McConaughy, B.L., Laird, C.B. and McCarthy, B.J. (1969) Biochemistry 8, 3289-3295.
13. Britten, R.J., Graham, D.E. and Neufield, B.R. (1974) Methods in Enzymol. 29, 263-270.
14. Southern, E.M. (1975) J. Mol. Biol. 98, 503-517.
15. Maxam, A.M. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74, 560-564.
16. Guo, L.-H. and Wu, R. (1982) Nucl. Acids Res. 10, 2065-2084.
17. Birnboim, H.C. and Doly, J. (1979) Nucl. Acids Res. 7, 1513-1523.
18. Aviv, H. and Leder, P. (1972) Proc. Natl. Acad. Sci. USA 69, 1408-1412.
19. Thomas, P.S. (1980) Proc. Natl. Acad. Sci. USA 77, 5201-5205.
20. Dickerson, R.E. (1971) J. Mol. Evol. 1, 26-45.
21. Gojobori, T., Li, W.-H. and Graur, D. (1982) J. Mol. Evol. 18, 360-369.
22. Perler, F., Efstratiades, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980) Cell 20, 555-566.
23. Chan, S. and Margoliash, E. (1966) J. Biol. Chem. 241, 507-515.
24. Lewin, B. (1980) Cell 22, 324-326.
25. Goldberg, M. (1979) Ph.D. Thesis, Stanford University, Palo Alto, California.
26. Kozak, M. (1978) Cell 15, 1109-1123.
27. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
28. Smith, M., Leung, D.W., Gillam, S. and Astell, C.R. (1979) Cell 16, 753-761.
29. Scarpulla, R.C., Agne, K.M. and Wu, R. (1982) Proc. Natl. Acad. Sci. USA 79, 739-743.
30. Scarpulla, R.C., Agne, K.M. and Wu, R. (1981) J. Biol. Chem. 256, 6480-6486.
31. Scarpulla, R.C. and Wu, R. (1983) Cell 32, 473-482.
32. Russell, P.R. and Hall, B.D. (1982) Mol. Cell Biol. 2, 106-116.
33. Boss, J.M., Darrow, M.D. and Zitomer, R.S. (1980) J. Biol. Chem. 255, 8623.
34. Guarente, L. and Mason, T. (1983) Cell 32, 1279-1286.
35. Matner, R. and Sherman, F. (1982) J. Biol. Chem. 257, 9811-9821.
36. Fyrberg, E.A., Bond, B.J., Hershey, N.D., Mixter, K.S. and Davidson, N. (1981) Cell 24, 107-116.
37. Jensen, E., Paludan, K., Hyldig-Nielsen, J., Jorgensen, P. and Marcker, K. (1981) Nature 291, 677-679.