Volume 10 Number 23 1982 Nucleic Acids Research

Independently evolving chicken histone H2B genes: identification of a ubiquitous H2B-specific 5' element

R.P.Harvey, A.J.Robins and J.R.E.Wells

Department of Biochemistry, University of Adelaide, Adelaide, South Australia 5001, Australia

Received 8 September 1982; Revised and Accepted 29 October 1982

ABSTRACT

The DNA sequence of two chicken histone H2B genes has been determined. Both genes code for the same H2B subtype. Except for conserved "promoter" elements, the sequences 5' to the protein coding regions are completely divergent, indicating that the genes are distantly related and are not evolving in concert. This presents an ideal situation for sequence comparisons. We have discovered a 13 bp, H2B specific homology block, 5' CTCATTTGCATAC 3' located close to the "TATA box". This motif is conserved in all H2B gene leader regions so far sequenced. One of the H2B genes is closely linked, in a divergent arrangement, to an H2A gene, and sequence data suggests that the linked genes share promoter elements.

INTRODUCTION

Preliminary characterization of vertebrate histone genes has shown that they are clustered, but highly disordered in arrangement (1). This is in distinct contrast to the regular organization of the sea urchin and Drosophila histone genes (1). Whether this structural disparity reflects major evolutionary changes in the nature of histone gene expression is at present speculative.

We have examined some structural details of histone genes in chickens in an attempt to broaden our understanding of histone gene regulation and evolution. Histone genes have already been exploited in the search for elements which regulate eukaryote genes. DNA sequence comparisons have been useful in locating important sequences both 5' and 3' to the genes (1,2). In addition, in vivo and in vitro transcription studies on native and manipulated histone genes have emphasised flanking regions which modulate expression (3-5). However, the relevance of these regions to general transcription, or to specific regulation has not yet been determined.

We report here the DNA sequences of two chicken H2B genes from the genomal clones λCH-01, λCH-05 and λCH-02 (6 and unpublished). Both of these

0305-1048/82/1023-7851 $2.00/0

genes code for the same H2B protein. We find that outside the protein coding regions, the untranslated and flanking sequences share few homologies. Significantly though, we have located a 13 bp sequence embedded in the 5' "promoter" region of the two chicken genes which is also highly conserved in H2B genes from other organisms. This sequence is, therefore, H2B gene-specific and likely to be important for the particular regulation of H2B genes.

MATERIALS AND METHODS

Enzymes and nucleotides: Restriction enzymes were purchased from New England Biolabs and conditions of digestion were as specified by their catalogue. E. coli DNA-polymerase "large fragment", T4 DNA ligase and T4 polynucleotide kinase were purchased from Boehringer. Calf intestinal phosphatase was purchased from Sigma and dialysed against 100 mM Tris-HCl pH 8, 10 μM $ZnSO_4$ before use. Conditions of incubation were 100 mM Tris-HCl pH 8, 0.1% SDS at 37°C (7).

Sequencing procedures: The H2B gene from λCH-01 was sequenced using the chemical methods of Maxam and Gilbert (8) except that 100% formic acid was used for the G + A reaction instead of pyridinium formate (R. Richards, personal communication). The 3' untranslated region and flanking sequences of this gene were determined from a clone which overlaps λCH-01 (λCH-05, unpublished) using the M13 phage-cloning and enzymatic sequencing procedure of Sanger et al. (9). The DNA sequence of the H2B gene from λCH-02 (7) was also determined using this method. Approximately 80% of DNA sequences were determined from both DNA strands. Where this was not the case, the sequence from one strand was determined several times.

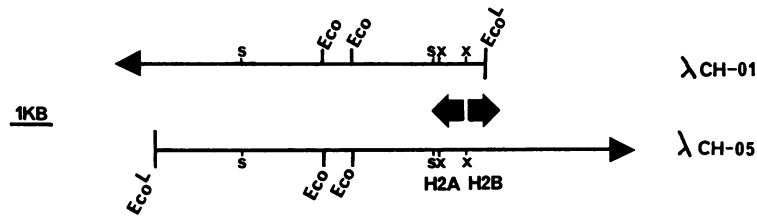
RESULTS AND DISCUSSION

The DNA sequence of the H2B gene contained within the chicken genomic clone, λCH-02 (6), is shown in Fig. 1.

The DNA sequence of an H2A/H2B gene pair contained within another non-overlapping chicken genomic clone, λCH-01 (6), is presented in Fig. 2. The insert of λCH-01 ends a few bases 3' to the H2B protein coding region. The 3' end of this gene was therefore sequenced from a genomic clone which overlaps λCH-01 (λCH-05: unpublished). The sequence of the H2A gene (contained within a 704 bp XhoI fragment, Fig. 2) has been previously reported (10), but its orientation relative to an adjacent (non-coding) SmaI/XhoI fragment was incorrectly determined, and so it is presented here in corrected form.

```
          -121                                                    -68
             CTGTTATCCAATCAGAGAGCAGATACAGAAGGCACTCGATTTGCATACTGCCCC

-67                                                               -1
  TATAAATAGGCGAGCAGTGCTCGCAGCCGGCACTCCGCTGCGCCGAAGGGATCGTGGAGAGTTCGAC

                                                                  51
 1ATG CCT GAG CCG GCC AAG TCC GCA CCC GCC CCC AAG AAG GGC TCC AAG AAG
     pro glu pro ala lys ser ala pro ala pro lys lys gly ser lys lys

 52                                                              102
  GCG GTC ACC AAG ACC CAG AAG AAG GGC GAC AAG AAG CGC AAG AAG AGC CGC
  ala val thr lys thr gln lys lys gly asp lys lys arg lys lys ser arg

 103                                                             153
  AAG GAG AGC TAC TCG ATC TAC GTG TAC AAG GTG CTG AAG CAG GTG CAC CCC
  lys glu ser tyr ser ile tyr val tyr lys val leu lys gln val his pro

 154                                                             204
  GAC ACG GGC ATC TCG TCC AAG GCC ATG GGC ATC ATG AAC TCG TTC GTC AAC
  asp thr gly ile ser ser lys ala met gly ile met asn ser phe val asn

 205                                                             255
  GAC ATC TTC GAG CGC ATC GCC GGC GAG GCG TCG CGC CTG GCG CAC TAC AAC
  asp ile phe glu arg ile ala gly glu ala ser arg leu ala his tyr asn

 256                                                             306
  AAG CGC TCG ACC ATC ACG TCG CGG GAG ATC CAG AXG GCC GTG CGG CTG CTG
  lys arg ser thr ile thr ser arg glu ile gln thr ala tyr arg leu leu

 307                                                             357
  CTG CCC GGC GAG CTG GCC AAG CAC GCG GTC TCC GAG GGC ACC AAG GCG GTC
  leu pro gly glu leu ala lys his ala val ser glu gly thr lys ala val

 358                                                             416
  ACC AAG TAC ACC AGC TCC AAG TAG AGCGGTGCGGATTACTCGATTTTAACCCAAAGGCT
  thr lys tyr thr ser ser lys

 417                                                             483
  CTTTTCAGAGCCACCATTTGTTCTAATAAAAGGGCTGTATTACTTTTTTTCTTTTTTTCCTGAGGGG

 484                                                             550
  TATAGCGTGGGTTAACTGAGTGAAATGGAAGCGAGTGCCTGAGGTATGTATATAATTGCTTAACTTC

 551                                                             617
  GCAGTTCGGAGGTCTCCGTTCCGAGTTAATTGAGCAGTAGCAACTCCCGGAGTTAACTGGGTTGGTC

 618                          645
  GGTAGCCGTGCTATTACTGCAGCACGGT
```

Figure 1.
The nucleotide sequence of the H2B gene from λCH-02 (6). Nucleotides 5' to the ATG initiation codon are negatively numbered. Only the anti-sense strand of the DNA is presented.




```
1GGGAGAAACTTCTTTTGGGAGAACGCTTTGGGCACGACTTTGTTAACGGAAGCATGGAAAGCGTTGCTATTATTACCCACCAAATAATACTGATAATAAATATGAGAAAA110
  CCCTCTTTGAAGAAAACCCTCTTGCGAAACCCGTGCTGAAACAATTGCCTTCGTACCTTTCGCAACGATAATAATGGGTGGTTTATTATGACTATTATTTATACTCTTTT

111AAAAAAGAAAAAAAAAAAAGCACGGCTCGAGTAATTCATAGACCTGCACGACTGTATTGCGGATTTCTGAGCTCTCCTGCGATGCTGTGGGTGGCTCTGAAAAGAGCTTTG220
  TTTTTTCTTTTTTTTTTTCGTGCCGAGCTCATTAAGTATCTGGACGTGCTGACATAACGCCTAAAGACTCGAGAGGACGCTACGACACCCACCGAGACTTTTCTCGAAAC

221GACTGTTCTCTCAGACAGCGCTGCCTCGCCGGTGC TCA CTT GGC CTT GGC CTT GTG GCT GTC GGT CTT CTT GGG CAG CAG CAC GGC CTG GAT312
  CTGACAAGAGAGTCTGTCGCGACGGAGCGGCCACG AGT GAA CCG GAA CCG GAA CAC CGA CAG CCA GAA GAA CCC GTC GTC GTG CCG GAC CTA
                                    STOP lys ala lys ala lys his ser asp thr lys lys pro leu leu val ala gln ile

313GTT GGG CAG CAC CCC GCC CTG CGC GAT GGT CAC CTT GCC CAG CAG CTT GTT GAG CTC CTC GTC GTT GCG GAT GGC CAG CTG CAG396
  CAA CCC GTC GTG GGG CGG GAC GCG CTA CCA GTG GAA CGG GTC GTC GAA CAA CTC GAG GAG CAG CAA CGC CTA CCG GTC GAC GTC
  asn pro leu val gly gly gln ala ile thr val lys gly leu leu lys asn leu glu glu asp asn arg ile ala leu gln leu

397GTG GCG GGG GAT GAT GCG CGT CTT CTT GTT GTC GCG GGC CGC GTT GCC CGC TAG CTC CAG GAT CTC GGC CGT CAG GTA CTC CAG480
  CAC CGC CCC CTA CTA CGC GCA GAA GAA CAA CAG CGC CCG GCG CAA CGG CGC ATC GAG GTC CTA GAG CCG GCA GTC CAT GAG GTC
  his arg pro ile ile arg thr lys lys asn asp arg ala ala asn gly ala leu glu leu ile glu ala thr leu tyr glu leu

481CAC GGC CGC TAG GTA CAC CGG GGC GCC GGC GCC CAC CCG CTC CGC GTA GTT GCC CTT GCG CAG CAG CCG GTG CAC GCG GCC CAC564
  GTG CCG GCG ATC CAT GTG GCC CCG CGG CCG CGG GTG GGC GAG GCG CAT CAA CGG GAA CGC GTC GTC GGC CAC CTG CGC CGG GTG
  val ala ala leu tyr val pro ala gly ala gly val arg glu ala tyr asn gly lys arg leu leu arg his val arg gly val

565GGG GAA CTG CAG CCC GGC GGC CGA CGA GCG CGA CTT GGC CTT GGC GCG CGC CTT CCC GCC CTG CTT TCC GCG CCC CGA CAT CGC648
  CCC CTT GAC GTC GGG CCG CCG GCT GCT CGC GCT GAA CCG GAA CCG CGC GCG GAA GGG CGG GAC GAA AGG CGC GGG GCT GTA GCG
  pro phe gln leu gly ala arg ser ser arg ser lys ala lys ala arg ala lys gly gly gln lys gly arg gly ser met

649AGCGACTGAACACTCAGAGAGCAAACAAGCCAGGAACGCAGCGCGCAACCTACGGAGCTCGGCTACGCTCTCGCTTTGGCTCCTCGCCTCTGCGCTGATAGGCTGACGTCT759
  TCGCTGACTTGTGAGTCTCTCGTTTGTTCGGTCCTTGCGTCGCGCGTTGGATGCCTCGAGCCGATGCGAGAGCCAAACCGAGGAGCGGAGACGCGACTATCCGACTGCAGA

760GGGTCCGCTCTGGGCAGCCAATGAGAAAGCGAATCGAATTTGAACCAATGAAAAGCAGTTATAAGGGAGGAAAGGCGTGTTCTCGAGTTCCGACCAATGAAAGAGTGCGAA870
  CCCAGGCGAGACCCGTCGGTTACTCTTTCGCTTAGCTTAAACTTGGTTACTTTTCGTCAATATTCCCTCCTTTCCGCACAAGAGCTCAAGGCTGGTTACTTTCTCACGCTT

871AGGAATGCTTCTCATTTGCATAGAGGGGCTATAAATAAATGCCTACGACCCCTTCGTTTCCATTCAGCGTCTCCTGGTCGTTTTTGTTCGCCTCGCTTCGTGAGCGCGTTG981
  TCCTTACGAAGAGTAAACGTATCTCCCCGATATTTATTTACGGATGCTGGGGAAGCAAAGGTAAGTCGCAGAGGACCAGCAAAAACAAGCGGAGCGAAGCACTCGCGCAAC

          met pro glu pro ala lys ser ala pro ala pro lys lys gly ser lys lys ala val thr lys thr gln lys lys gly
982TGCCACT ATG CCC GAG CCG GCT AAG TCC GCG CCC GCC CCG AAG AAG GGC TCT AAG AAG GCG GTC ACC AAG ACC CAG AAG AAG GGC1069
  ACGGTGA TAC GGG CTC GGC CGA TTC AGG CGC GGG CGG GGC TTC TTC CCG AGA TTC TTC CGC CAG TGG TTC TGG GTC TTC TTC CCG

  asp lys lys arg lys lys ser arg lys glu ser tyr ser ile tyr val tyr lys val leu lys gln val his pro asp thr gly
106/GAC AAG AAG CGC AAG AAG AGC CGC AAG GAG AGC TAC TCG ATC TAC GTG TAC AAG GTG CTG AAG CAG CTG CAC CCC GAC ACG GGC1150
  CTG TTC TTC GCG TTC TTC TCG GCG TTC CTC TCG ATG AGC TAG ATG CAC ATG TTC CAC GAC TTC GTC GAC GTG GGG CTG TGC CCG

  ile ser ser lys ala met gly ile met asn ser phe val asn asp ile phe glu arg ile ala gly glu ala ser arg leu ala
1151ATC TCG TCC AAG GCC ATG GGC ATC ATG AAC TCG TTC GTC AAC GAC ATC TTC GAG CGC ATC GCC GGC GAG GCG TCG CGC CTG GCG1234
  TAG AGC AGG TTC CGG TAC CCG TAG TAC TTG AGC AAG CAG TTG CTG TAG AAG CTC GCG TAG CGG CCG CTC CGC AGC GCG GAC CGC

  his tyr asn lys arg ser thr ile thr ser arg glu ile gln thr ala val arg leu leu leu pro gly glu leu ala lys his
1235CAC TAC AAC AAG CGC TCG ACC ATC ACG TCG CGG GAG ATC CAG ACA GCC GTG CGG CTG CTG CTG CCC GGC GAG CTG GCC AAG CAC1318
  GTG ATG TTG TTC GCG AGC TGG TAG TGC AGC GCC CTC TAG GTC TGT CGG CAC GCC GAC GAC GAC GGG CCG CTC GAC CGG TTC GTG

  ala val ser glu gly thr lys ala val thr lys tyr thr ser ser lys STOP
1319GCG GTC TCG GAG GGC ACC AAG GCG GTC ACC AAG TAC ACC AGC TCC AAG TAA GCTGTTTTCATCCCCCGTGCCAGAGTCGCTCGACCCAAAGGCT1402
  CGC CAG AGC CTC CCG TGG TTC CGC CAG TGG TTC ATG TGG TCG AGG TTC ATT CGACAAAAGTAGGGGGCACGGTCTCAGCGAGCTGGGTTTCCGA

1403TCAGAGCCACCCACCTTGCCAGAGAAAGAGCTGTAGTCATATCAGAAGAAAATAGCATAAGTTAATCCGGCTTTCTACTTCACATCTTTAAATACTTTTCTAAAAAA1513
  GAAAAGTCTCGGTGGGTGGAACGGTCTCTTTCTCGACATCAGTATAGTCTTCTTTTATCGTATTCAATTAGGCCGAAAGATGAAGTGTAGAAATTTATGAAAAGATTTTTT

1514AAATTCCTGTAAATACTATTACCTGCCAGCTTTGTTCCAGAAAAT1558
  TTTAAGGACATTTATGATAATGGACGGTCGAAACAAGGTCTTTTA
```

Figure 2.

The nucleotide sequence of the H2A/H2B gene pair contained within the genomal clones λCH-01 (6) and λCH-05 (unpublished). Nucleotides are

numbered continuously from the 3' trailer region of the H2A gene (λCH-01) through to the 3' trailer region of the H2B gene (λCH-05). Both strands of the DNA are presented as the genes are in divergent orientation. Amino acids are aligned next to the anti-sense strand of the duplex. A restriction map is shown to indicate the overlapping areas of λCH-01 and λCH-05. The position and orientation of the H2A and H2B genes are shown by the large arrows. S = SmaI, X = XhoI, Eco = EcoRI, Eco^{L} = EcoRI linker.

The H2B Protein Sequence

The proteins encoded in the two H2B genes (Figs. 1 and 2) are identical. This is the first report of a complete chicken H2B amino acid sequence. Urban et al. (11) have isolated the major core histones expressed in chicken erythrocytes and, by peptide mapping, have determined a few of the sites which are substituted in variants of the same class. The chicken H2B sequence corresponds to the H2B.1 variant, distinguishing it from a less predominant erythrocyte subtype, H2B.2. It also corresponds exactly to some partial amino acid sequences determined by Van Helden et al. (12) from chicken erythrocyte H2B.

A comparison of the chicken H2B protein sequence with H2B protein sequences from other species is shown in Fig. 3. The evolution of the H2B protein has been discussed in detail by Von Holt et al. (13). The C-terminal 89 amino acids are very highly conserved across wide species barriers. This region abutts an extremely basic area of nine amino acids, of which 7-8 are arginine or lysine, and the N-terminal region which is rapidly evolving.

The chicken H2B sequence is consistent with the observation that the H2B protein is an evolutionary hybrid. All non-conservative changes (except one) between the chicken, calf and trout sequences are in the rapidly evolving N-terminal region. The only amino acid in the chicken sequence that has not been observed in H2B sequences of other species is the lysine at position 31. All other species from which sequences have been determined (13) have arginine in this position, but the change is conservative and embedded in the highly basic domain.

H2B mRNA Sequence

The DNA sequences of the two H2B genes differ in only seven positions within the protein coding region. The changes are to isocoding triplets, resulting in an identical coding capacity.

The 5' and 3' untranslated regions of the two genes are totally divergent (except for the "CAP" sequence and the ubiquitously conserved histone terminator region, Fig. 4). In addition, the sequence data suggests

```
H2B
(1) CHICK                                                          PEPAKSAPAPKKGSKK
(2) CALF THYMUS AND HUMAN SPLEEN
(3) TROUT TESTIS                                                             --
(4) PATELLA TESTIS                                                      PPKVSSKGAKK
(5) DROSOPHILA                                                          PPKTSGKAAKKA
(6) SEA URCHIN VARIANT B                PRSPAKTSPRKGSPRKGSPRKGSPSRKASPKRGGKGAKR

                    *
(1) AVTKTQKKGDKKRKKSRKESYSIYVYKVLKQVHPDTGISSKAMGIMNSFVNDIFERIAGEASRLAHYNKRST
(2)     A   DG    R       V                                              S
(3)      AG  G    R       A                V
(4) AGKAKAARS     RR                              S                A
(5) GKAQKNITKT  K RK      A                       S                A
(6) AGKGGRRRRVV  RRRR     G I                  R  SV       V      S     TSA R

(1) ITSREIQTAVRLLLPGELAKHAVSEGTKAVTKYTSSK
(2)
(3)
(4)
(5)
(6) V                                   T R
```

Figure 3.

Comparison of the chicken H2B amino acid sequence predicted by the genes shown in Fig. 1 and Fig. 2 with H2B sequences of other species (2). Single letter amino acid abbreviations are used: A - Ala, C - Cys, D - Asp, E - Glu, F - Phe, G - Gly, H - His, I - Ile, K - Lys, L - Leu, M - Met, N - Asn, P - Pro, Q - Gln, R - Arg, S - Ser, T - Thr, V - Val, Y - Tyr. The chicken sequence is shown in full. For the other species, amino acids are shown only when different to the chicken sequence.

that the mRNAs are of different lengths. If the length of the transcript is taken from the A in the CAP initiation sequence (5), to the A within the conserved terminator sequence (5), the λCH-02 H2B gene and the λCH-01/ λCH-05 H2B gene would produce transcripts of 465 bps and 494 bps respectively.

In the human globin system, the relatedness of untranslated sequences reflects the evolutionary relatedness of the respective genes (14). The fact that extensive sequence divergence has occurred in the untranslated regions of two copies of a chicken H2B gene, argues that the two genes are very distantly related and that infrequent correctional events (15) have occurred between the genes. This supports, at a DNA sequence level, the idea that infrequent correctional events have allowed "randomization" of the vertebrate histone gene arrangement (6,16). Similarly, the distinct contrast between the divergence seen in these chicken genes, and the conserved nature of histone gene copies within the sea urchin and Drosophila systems (17) lends support to the notion that the invertebrate histone gene clusters are highly conserved because of mechanisms of concerted evolution (15), and not because of specific selection pressures (6,16).

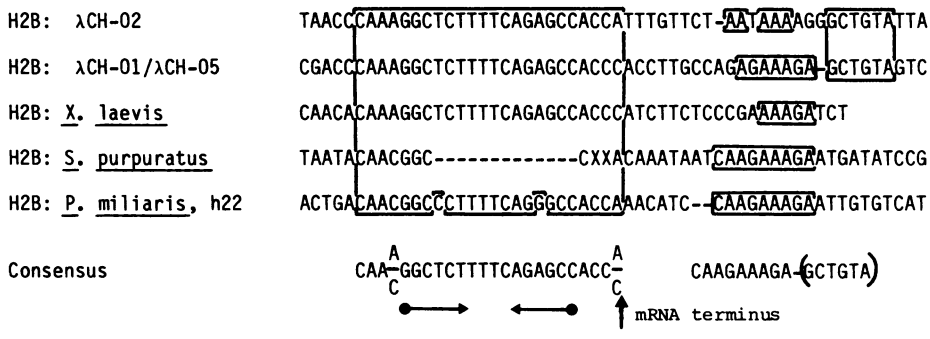


Figure 4.

Comparison of the DNA sequences from the 3' untranslated regions, and 3' flanking regions of the chicken H2B genes and H2B genes from X. laevis (21), S. purpuratus (24) and P. miliaris (5). Consensus sequences are indicated. The mRNA dyad-symmetry consensus sequence is derived from the H2B genes indicated. The convergent arrows indicate the homologous nucleotides in the stem of the hyphenated dyad-symmetry element. The consensus sequence in the 3' flanking region consists of the sea urchin consensus sequence (1) and in addition the homology block observed in the chicken H2B genes (bracketted, see text). Homologies in the H2B gene sequences to the indicated consensus sequences are blocked in.

Termination of histone gene transcription seems to be dependent on a specific sequence in the 3' untranslated region, which contains a dyad symmetry element (18). This sequence has been conserved in all histone genes so far sequenced except for yeast histone genes (18), the avian erythroid-specific H5 gene (19), and an unusual chicken H3 gene which contains intervening sequences (20). The terminator sequence is also conserved in these chicken H2B genes (Fig. 4) S_1-mapping data, and sequencing of histone mRNAs and cDNA clones places the histone mRNA terminus at the sequence 5' CCACCA-OH 3' (5). One chicken H2B gene, together with a Xenopus gene (21) (Fig. 4), contain the sequence 5' CCACCC 3' in this position. These mRNAs may, therefore, end with a C residue.

H2B 5' Leader Sequences

Comparison of DNA sequence of genes from a variety of species has been useful in defining sequences important for general polymerase II transcription (5,14). However, apart from the "TATA box" (selector sequence), these sequences are mostly "modulators" in that they influence the rate of transcription. Their manipulation in surrogate genetic systems results in either a stimulation (3), or a depression (3,4,22) in the rate of transcription initiation. Few sequences, on the other hand, have been

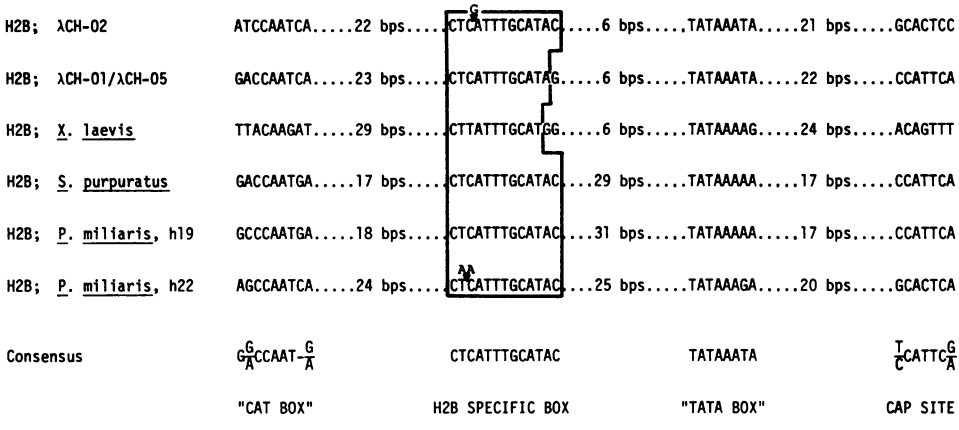


Figure 5.
5' Leader region homologies in H2B genes. Regions of 5' homology between the chicken H2B genes and H2B genes from X. laevis (21), S. purpuratus (24) and P. miliaris (2) are tabulated. Consensus sequences for the "CAT box", "TATA box" and "CAP" sequence are those of Hentschel and Birnstiel (1) derived from a comparison of histone genes from various species. Homology to the H2B-specific box is blocked in. The distances between homology blocks are indicated in base pairs (bps).

implicated in the specific switching of developmentally regulated genes. This may be because these sequences exist at a greater distance from coding regions than anticipated. The β-globin gene family, for instance, shows considerable cross-species homology in extensive 5' regions (23). Alternatively, this aspect of regulation may be influenced more by chromatin architecture and DNA or histone modifications.

The divergent nature of the untranslated regions of the two H2B genes reported in this paper suggests that they are not evolving in concert. This presents an ideal situation for sequence comparisons - two distantly related genes with identical coding capacity, evolving essentially independently within the same species. We have compared the 5' leader regions of the two chicken H2B genes and of H2B genes from X. laevis (21), S. purpuratus (24), and P. miliaris (2). The regions of sequence found to be conserved are shown in Fig. 5.

Three conserved sequence elements are apparent. The "TATA box" (selector sequence) and the "CAT box" (modulator sequence) are present in all H2B genes compared. These motifs are shared by many other eukaryote genes and are likely to be constituents of the general polymerase II promoter.

A third element, consisting of a tightly conserved 13 bp sequence, is present in the H2B genes of all of the species. This sequence has not been found in the prelude regions of other histone genes, nor other polymerase II genes. It, therefore, represents an element specific to H2B genes and conserved rigidly between diverse species. The position of this element relative to the "TATA box" is constant within the vertebrate genes (6 bps) and relatively constant within the sea urchin genes (25-31 bps). The location of the H2B-specific sequence relative to the other general promoter elements strongly suggests that it participates in the specific regulation of H2B gene transcription. It will be a particularly interesting sequence to manipulate in surrogate genetic systems.

H2B 3' Trailing Sequences

A comparison of 3' trailer sequences of the chicken H2B genes shows interesting features. A homology block (9 bps) close to the conserved mRNA termination sequence which is remarkably conserved in sea urchins, but only weakly conserved in other species (5), is also present in recognisable form in the chicken H2B genes (Fig. 4). A similar weak homology is also seen in the X. laevis H2B 3' sequence. The extent of the homology is considerably less than perfect (5/9 for λCH-02 H2B gene; 7/9 for λCH-05 H2B gene; 6/9 for X. laevis H2B gene), however, the maintenance of this feature in vertebrate histone genes must be considered to be significant because it remains in the chicken H2B genes despite maximal divergence in other flanking sequences.

This 3' flanking homology block may have a function in histone mRNA termination or maturation. It is known that the conserved dyad symmetry element present in histone mRNAs is not alone sufficient to terminate histone gene transcription, and that other sequences in the 3' flanking region are also required (18). Although it has not been investigated by deletion mapping, the second homology block in the 3' flanking region may compliment the dyad symmetry element in performing the termination function. It is puzzling, however, that in vertebrates the mRNA dyad symmetry element is so well conserved, whereas the 3' flanking homology block exhibits considerable variation. In sea urchins, this second homology block is almost invariant. One possible reason for this is that in chicken the second homology block is more extensive (Fig. 6), containing an invariant GCT in all chicken histone genes for which we have sequencing data (except for the H5 gene and the unusual H3 gene (20)). In the two chicken H2B genes, this third homology consists of a 6 bp extension (GCTGTA) to the second homology block. The other chicken genes show variable agreement with

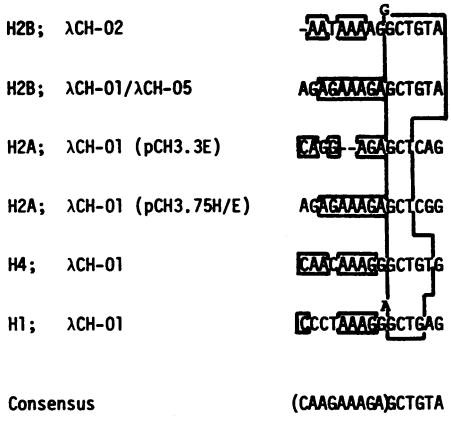


Figure 6.
Comparison of the 3' flanking histone homology regions in chicken histone genes. All genes are derived from the genomal clones λCH-01 and λCH-02 (6). The H2B gene of λCH-01 was cut short (in the 3' untranslated region) by the cloning procedure, but is also present in an overlapping clone, λCH-05 (unpublished). A consensus sequence is indicated (see Figure 4). Homology to this consensus sequence is blocked in.

this sequence, but in all cases the GCT is present, and always abutts the second homology block.

The extended homology is not obvious in the sea urchin or Xenopus gene sequences available and so may represent a chicken specific sequence. If this extended sequence does function in histone gene transcription termination, then at least one aspect of the termination machinery has evolved to recognize a different (but related) DNA sequence in chickens. The differential evolution of the mRNA dyad symmetry element, and the 3' flanking homology block implicates a multi-component termination mechanism as in bacterial systems (25).

Another noteworthy feature of the chicken 3' flanking regions is the stretch of 14 T residues broken centrally by a C residue, close to the λCH-02 H2B gene (Fig. 1). An identical sequence occurs downstream from the λCH-01 H2A gene but it is not present in the closely linked λCH-01/λCH-05 H2B gene (Fig. 2). Interestingly, the same sequence in reverse orientation is present in the 3' region of the unusual chicken H3 gene isolated by Engel et al. (20). The function of such features is at present unknown.

The Divergent H2A/H2B Gene Pair

The H2A and H2B genes of λCH-01 (Fig. 2) are tightly linked and arranged in a divergent fashion. This type of arrangement was first observed for histone genes in the Drosophila system (26), and because transcription occurred from both DNA strands, precluded the existence of a single poly-cistronic histone transcript. Since then, histone, and other eukaryote genes, have been observed in non-tandem arrangements (1,27). In

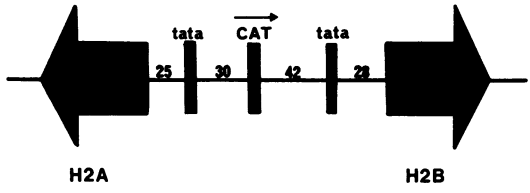


Figure 7.
The intergene region of the λCH-01 H2A/H2B gene pair. Putative promoter elements ("CAT box" and "TATA box") are blocked in to indicate the symmetry of their arrangement. The distance between these elements is indicated in base pairs. The genes are represented by large arrows, the direction of which indicates the transcriptional orientation (5' → 3').

some cases the tight linkage of divergently arranged gene pairs implies that some expressional advantage is to be found in such an arrangement. For instance, approximately 300 base pairs separates the histone H3/H4 gene pair of Drosophila (26) and 264-325 base pairs separates the silk moth chorion A/B gene pairs (27). It has been proposed that the divergent arrangment of tightly linked gene pairs may facilitate co-ordinate expression of those genes by the sharing of promoter elements (26,27). In support of this, chorion A/B gene pairs are co-ordinately expressed in particular developmental periods. DNA sequencing data on silk moth chorion genes has, however, failed to reveal any features which would support the notion of shared promoter elements (27). Each paired gene has its own "TATA box" and possible "modulator" sequences, and no strong symmetry elements exist in the intergenic regions.

The chicken H2A and H2B genes (Fig. 2) are tightly linked indeed. Only 135 base pairs separates the proposed transcription initiation sites (CAP sites, Fig. 5 and Ref. 10) of the two genes, and only 77 bps separates their "TATA boxes". It seemed likely that the promixity of the two genes would result in an overlapping of "promoter elements" with the possibility that some aspects of the promoter would be shared. A computer analysis did not reveal any strong dyad symmetry within the intergenic region. However, the "CAT box" (CCAAT) for the H2B gene lies almost centrally between the "TATA boxes" of the two genes (Fig. 7). No true "CAT" sequence can be found for the H2A gene. This striking symmetry in the location of promoter elements seems to suggest that the H2B "CAT" sequence may be shared by both genes. It will be particularly instructive to manipulate this intergenic region in in vivo or in vitro transcription systems.

Although such an arrangement might facilitate co-ordinate expression

of the gene pair, it is not necessarily implied that transcription of each gene would occur at the same rate. Mutation of the "CAT box" region of Herpes Simplex Virus thymidine kinase gene (S. McKnight, personal communication) can result in an increase in transcriptional efficiency (the "CAT box" has a modulator function on the efficiency of transcription initiation).

In the chicken system, if the H2B "CAT site" serves to modulate both H2A and H2B transcription, then the H2A gene would have an inverted "CAT" sequence and may, therefore, possess a modulator of different activity.

ACKNOWLEDGMENTS

We thank P. Krieg, L. Coles and R. D'Andrea for releasing sequence and mapping data prior to publication, A. Moorman for the sequences of Xenopus histone genes while in press, R. Sturm for some M13 cloning assistance and R. Symons for radionucleotides.

BIBLIOGRAPHY

1. Hentschel, C.C. and Birnstiel, M.L. (1981). Cell 25, 301-313.
2. Busslinger, M., Portmann, R., Irminger, J.C. and Birnstiel, M.L. (1980). Nucleic Acids Res. 8, 957-977.
3. Grosschedl, R. and Birnstiel, M.L. (1980). Proc. Natl. Acad. Sci. USA. 77, 7102-7106.
4. Grosschedl, R. and Birnstiel, M.L. (1982). Proc. Natl. Acad. Sci. USA. 79, 297-301.
5. Hentschel, C.C., Irminger, J.C., Bucher, P. and Birnstiel, M.L. (1980). Nature 285, 147-151.
6. Harvey, R.P., Krieg, P.A., Robins, A.J., Coles, L.S. and Wells, J.R.E. (1981). Nature 294, 49-53.
7. Shinagawa, M. and Padmanabhan, R. (1979). Analytical Biochemistry 95, 458-464.
8. Maxam, A.M. and Gilbert, W. (1980). Methods Enzym. 65, 499-560.
9. Sanger, F., Niclen, S. and Coulson, A.R. (1977). Proc. Natl. Acad. Sci. USA. 74, 5463-5467.
10. D'Andrea, R., Harvey, R.P. and Wells, J.R.E. (1981). Nucleic Acids Res. 9, 3119-3128.
11. Urban, M.K., Franklin, S.G. and Zweidler, A. (1979). 18, 3952-3959.
12. Van Heldon, P., Strickland, W.N., Brandt, W.F. and Von Holt, C. (1978). Biochim. Biophys. Acta 533, 278-281.
13. Von Holt, C., Strickland, W.N., Brandt, W.F. and Strickland, M.S. (1979). FEBS Letters 100, 201-218.
14. Efstratiadis, A., Psakony, J.W., Maniatis, T., Lawn, R.L., O'Connell, C.O., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980). Cell 21, 653-668.
15. Baltimore, D. (1981). Cell 24, 592-594.
16. Heintz, N., Zernik, M. and Roeder, R.G. (1981). Cell 24, 661-668.
17. Kedes, L.H. (1979). Ann. Rev. Biochem. 48, 837-870.
18. Birchmeier, C., Grosschedl, R. and Birnstiel, M.L. (1982). Cell 28, 739-745.

19. Krieg, P.A., Robins, A.J., Colman, A. and Wells, J.R.E. Submitted for publication.
20. Engel, J.D., Sugarman, B.J. and Dodgson, J.B. (1982). Nature 297, 434-436.
21. Moorman, A.F.M., Boer, P.A.J., De Laff, R.T.M. and Destree, O.H.J. (1982). FEBS Letters, in press.
22. McKnight, S.L., Gavis, E.R., Kingsbury, R. and Axel, R. (1981). Cell 25, 385-398.
23. Moschonas, N., de Boer, E. and Flavell, R.A. (1982). Nucleic Acids Res. 10, 2109-2134.
24. Sures, I., Lowry, J. and Kedes, L.H. (1978). Cell 15, 1033-1044.
25. Platt, T. (1981). Cell 24, 10-23.
26. Lifton, R.P., Goldberg, M.L., Karp, R.W. and Hogness, D.S. (1977). Cold Spring Harbor Symp. Quant. Biol. 42, 1047-1051.
27. Jones, C.W. and Kafatos, F.C. (1980). Cell 22, 855-867.
28. Keene, M.A., Corces, V., Lowenhaupt, K. and Elgin, S.C.R. (1981). Proc. Natl. Acad. Sci. USA. 78, 143-146.
29. Link, G. and Bogorad, L. (1980). Proc. Natl. Acad. Sci. USA. 77, 1832-1835.