Vertebrate histone genes: nucleotide sequence of a chicken H2A gene and regulatory flanking sequences

Richard D'Andrea, Richard Harvey and Julian R.E.Wells

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

Received 27 April 1981

ABSTRACT

The DNA sequence of a chicken genomal fragment containing a histone H2A gene has been determined. It contains extensive 5' and 3' flanking regions and encodes a protein identical in sequence to the histone H2A protein isolated from chicken erythrocytes[1].

In the 5' flanking region, a possible "TATA box" and three possible "cap sites" can be recognised upstream from the initiation codon. To the 5' side of the "TATA box" is found an unusual sequence of 21 A's interrupted by a central G residue. It occupies the same relative position as the P. miliaris H2A gene-specific 5' dyad symmetry sequence and the "CCAAT box" seen in other eukaryotic polymerase II genes but is clearly different from both.

A significant feature of the 3' non-coding region is the presence of a 23 base-pair sequence that is nearly identical to a conserved region found in sea urchin histone genes[2].

The coding region is extremely GC rich, with strong selection for these bases in the third position of codons. Not a single coding triplet ends in U. No intervening sequences were found in this gene.

INTRODUCTION

The histones comprise a set of five small basic proteins involved in the packaging of eukaryotic chromosomes. These proteins show a remarkable degree of sequence conservation and in the two most studied invertebrate systems, sea urchin and Drosophila, their genes are organised in tandemly repeated units[3]. Different histone subtypes are expressed in a tissue-specific manner. These variants together with post-translational modifications of histones may clearly contribute to variations of nucleosome structure and chromatin conformation.

Recently, we have found that chicken histone genes are not organised in a simple compact repeat[4,5] and thus do not conform to arrangements seen in invertebrates. We report here the nucleotide sequence of one of the two H2A genes contained within a recombinant clone, λCH-01[4], with particular reference to putative control regions on both the 5' and 3' sides of the coding region.

MATERIALS AND METHODS

Subclones Containing H2A Gene Sequences

A 3.3 Kb EcoRI fragment containing H2A coding sequences[4] was subcloned into the EcoRI site of the plasmid pBR325 using standard procedures. This clone is designated pCH3.3E.

Further restriction analysis in conjunction with the known amino acid sequence of chicken histone H2A and limited nucleotide sequence data[4] indicated that XhoI cleaved the 3.3 Kb insert of pCH3.3E to generate a 0.7 Kb fragment containing the complete gene and 5' and 3' non-coding regions (assuming no intervening sequences). This 0.7 Kb fragment was also subcloned (S. Fields, Cambridge, U.K.) into the SalI sites of the Vector M13mp7, and M13 recombinants containing the insert in opposite orientations were selected by mini-screening.

DNA Sequencing

For the chemical degradation procedure all strategies were based on the methods of Maxam and Gilbert[6].

We found that DNA fragments eluted from preparative acrylamide gels failed to digest adequately with restriction enzymes. When thin (0.5 mm) preparative gels were utilised there was no difficulty. Excellent resolution was obtained and 50-100 µg of DNA loaded in 8 cm slots could be accommodated.

The dideoxy chain termination reactions[7] were also carried out on M13 cloned material using a synthetic primer[8]. There was substantial agreement of data from the two sequencing procedures. Ambiguities were resolved by further sequence analysis.

RESULTS AND DISCUSSION

The H2A gene is located within a region of pCH3.3E bordered by XhoI restriction sites. The position of this H2A gene in the original clone, λCH-01, and the subclone, pCH3.3E, are shown in Figure 1. The position of "spacer" and another coding region within pCH3.3E (now known to be the H2B coding region) are shown together with the direction of transcription of the H2A gene. Since there are no sites for XhoI in pBR325 the 0.7 Kb H2A gene-specific fragment bounded by these sites (Fig. 1) was easily purified from XhoI digested pCH3.3E by sucrose gradient centrifugation. Confirmation of sequences was obtained by repeated analysis and with reference to sequences obtained by dideoxy methodology (S. Fields). There was particularly good agreement of sequence data from these two methods in the 5' and 3' ends of

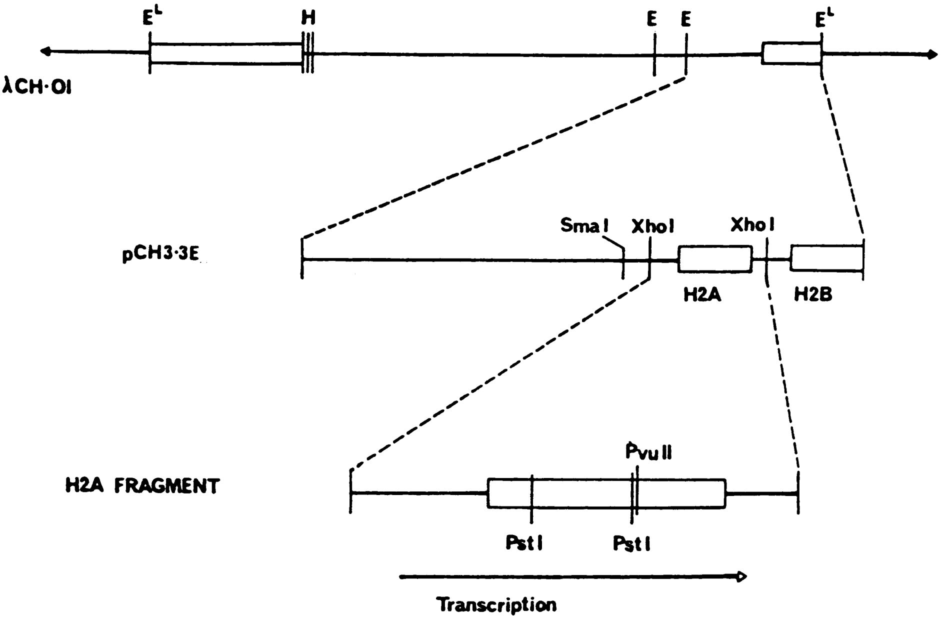


Figure 1. The coding regions within the 14 Kb insert of λCH-01 (indicated by boxes) were detected with chicken histone cDNA[4]. The 3.3 Kb EcoRI fragment was subcloned into pBR325. The recombinant plasmid pCH3.3E contains convenient XhoI sites flanking the H2A gene.

E	-	EcoRI
E^L	-	EcoRI linker
H	-	HindIII.

the fragment. Towards the centre, dideoxy ladders were more difficult to read with accuracy. This region was well covered by the chemical degradation method.

The SmaI/XhoI fragment adjacent to the 5' side of the XhoI/XhoI H2A coding fragment (Fig. 1) was also sequenced by chemical degradation methods to extend the region in which presumptive 5' control regions might occur.

The complete sequence of the H2A gene-specific fragment and 5' flanking sequences are shown in Figure 2.

Codon Usage

The nucleotide sequence of the histone H2A coding region has an extremely high GC content (69% compared with 41% for the genome[9]). This is particularly evident in the third base positions of codons. The preference for codons ending in G and C is stronger than that observed in the sea

```
                                                      -333      -326
                                                   5'    GGGAGAAA

-325                                                            -261
  CTTATTTTGGGAGAACGCTTTGGGCACGACTTTGTTAACGGAAGCATGGAAAGCGTTGCTATTAT

-260                                                            -196
  TACCCACCAAATAATACTGATAATAAATATGAGAAAAAAAAAAGAAAAAAAAAAAGCACGGCTCG

-195                                                            -131
  AGAACACGCCTTTCCTCCCTTATAACTGCTTTTCATTGGTTCAAATTCGATTCGCTTTCTCATTG

-130                                                             -66
  GCTGCCCAGAGCGGACCCAGACGTCAGCCCATCAGCGCAGAGGCGAGGAGCCAAAGCGAGAGCGT

 -65                                                              -1
  AGCGAGCCCGTAGGTTGCGCGCTGCGTTCTTGGCTTGTTTGCTCTCTGAGTGTTCAGTCGCTGCG

 1                                                                51
  ATG TCG GGG CGC GGA AAG CAG GGC GGG AAG GCG CGC GCC AAG GCC AAG TCG
  met ser gly arg gly lys gln gly gly lys ala arg ala lys ala lys ser

 52                                                              102
  CGC TCG TCG CGG GCC GGG CTG CAG TTC CCC GTG GGC CGC GTG CAC CGG CTG
  arg ser ser arg ala gly leu gln phe pro val gly arg val his arg leu

103                                                              153
  CTG CGC AAG GGC AAC TAC GCG GAG CGG GTG GGC GCC GGC GCC CCG CTG TAC
  leu arg lys gly asn tyr ala glu arg val gly ala gly ala pro val tyr

154                                                              204
  CTA GCG GCC GTG CTG GAG TAC CTG ACG GCC GAG ATC CTG GAG CTA GCG GGC
  leu ala ala val leu glu tyr leu thr ala glu ile leu glu leu ala gly

205                                                              255
  AAC GCG GCC CGC GAC AAC AAG AAG ACG CGC ATC ATC CCC CGC CAC CTG CAG
  asn ala ala arg asp asn lys lys thr arg ile ile pro arg his leu gln

256                                                              306
  CTG GCC ATC CGC AAC GAC GAG GAG CTC AAC AAG CTG CTG GGC AAG GTG ACC
  leu ala ile arg asn asp glu glu leu asn lys leu leu gly lys val thr

307                                                              357
  ATC GCG CAG GGC GGG GTG CTG CCC AAC ATC CAG GCC GTG CTG CTG CCC AAG
  ile ala gln gly gly val leu pro asn ile gln ala val leu leu pro lys

358                                                              412
  AAG ACC GAC AGC CAC. AAG GCC AAG GCC AAG TGA GCACCGGCGAGGCAGCGCTGTC
  lys thr asp ser his lys ala lys ala lys STOP

413                                                              478
  TGAGAGAACAGTCCAAAGCTCTTTTCAGAGCCACCCACAGCATCGCAGGAGAGCTCAGAAATCCGC

479                             510
  AATACAGTCGTGCAGGTCTATGAATTACTCGA   3'
```

Figure 2. The nucleotide sequence derived from analysis of the H2A specific Xho-XhoI fragment and the adjacent XhoI-SmaI fragment (Fig. 1). In the 5' untranslated region a long A stretch, probable "TATA box" and possible "cap sites" are underlined. In the 3' untranslated region a conserved sequence with dyad symmetry is also marked. The predicted amino acid sequence agrees with the data of Laine et al.[1]. All bases are numbered relative to the ^{1}AUG.

urchin histone genes[3] or chicken α and β-globin mRNAs[10,11] (see Table 1). Selection in the mRNA sequence for such a high GC content may result in a more stable overall secondary structure. Remarkably there is not a single U as the third base of any chicken H2A gene codon despite the fact that of 128 codons only 25 are precluded from having U in this position.

The 5' Flanking Region

DNA sequence analysis of a number of histone genes from sea urchin species has revealed conserved sequences in the 5' region[12]. These sequences have also been found, although modified, upstream from other eukaryotic genes and it has been suggested that these sequences are shared by genes transcribed by polymerase II.

A comparison of the 5' sequence of the chicken histone H2A gene with the sea urchin H2A 5' sequences and the prelude sequence of the chicken ovalbumin gene has revealed sequences related to those found by Busslinger *et al.*[12].

An AT-rich region, possibly representing the "TATA box" is found 175 nucleotides to the 5' side of the initiation codon of the chicken H2A gene (Fig. 3a). Examination of a number of gene sequences shows that the "TATA box" is located predominantly with the first "T" 31 nucleotides (range, 27-33) from the initiation of transcription[13]. It has been shown that the sea urchin histone mRNA 5' termini map in, or very near to, the "cap sequence", 5' PyCATTCPu 3', found in all sea urchin prelude sequences[14].

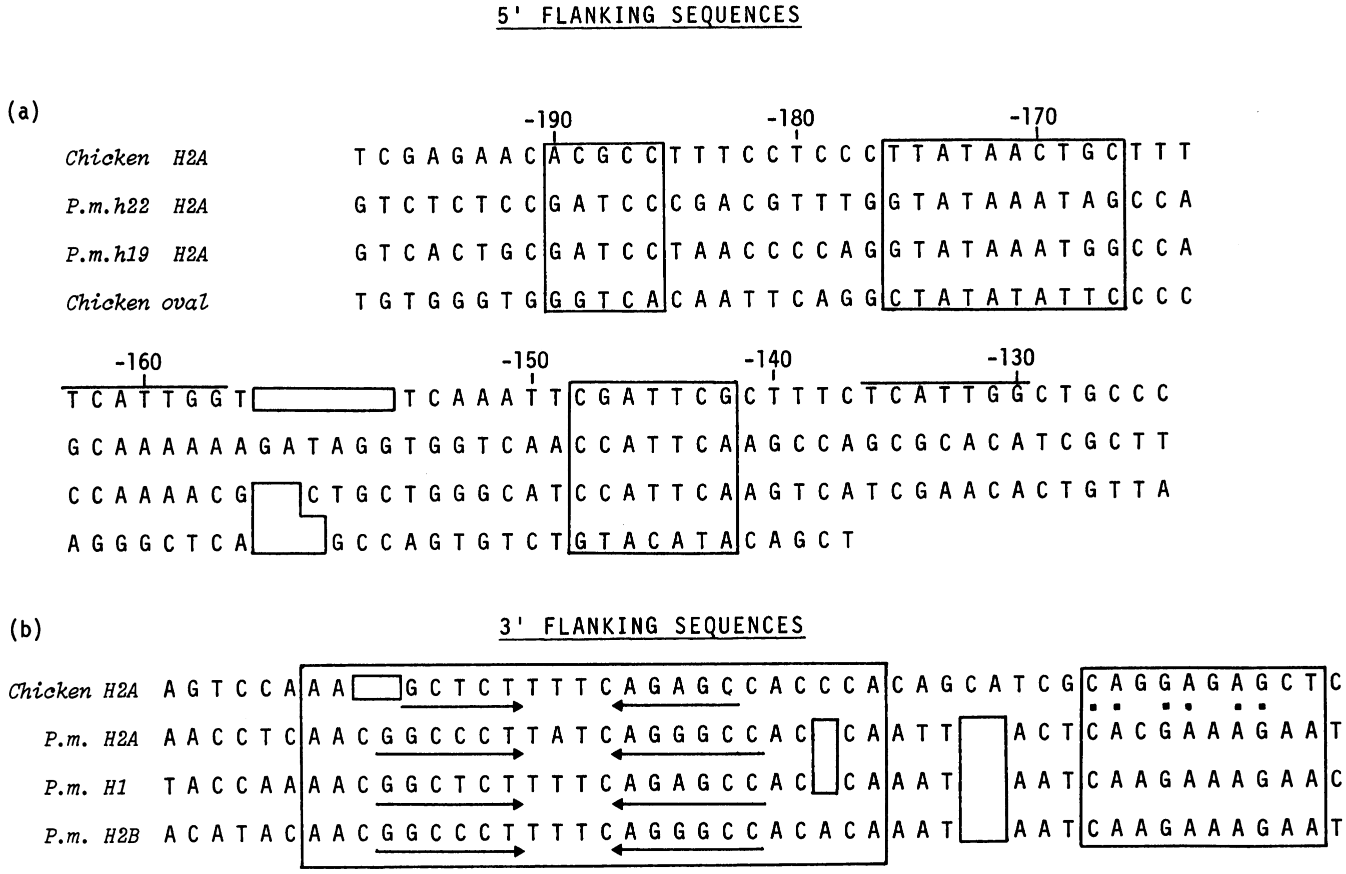
In the chicken H2A sequence there are three possible cap sites

TABLE 1

PREFERENCE FOR CODONS ENDING IN G AND C IN THE CHICKEN H2A GENE

	NNG	NNC
S. purpuratus Histone genes	24%	37%
P. miliaris Histone genes	27%	35%
Chicken β-globin mRNA	30%	49%
Chicken α_S-globin mRNA	27%	54%
Chicken Histone H2A gene	52%	46%

The table shows a comparison of third base codon usage in the chicken H2A gene, in sea urchin histone genes[3] and chicken globin genes[10,11].


5' FLANKING SEQUENCES
(a)
-190
-180
-170
Chicken H2A TCGAGAACACGCCTTTCCTCCCTTATAACTGCTTT
P.m.h22 H2A GTCTCTCCGATCCCGACGTTTGGTATAAATAGCCA
P.m.h19 H2A GTCACTGCGATCCTAACCCCAGGTATAAATGGCCA
Chicken oval TGTGGGTGGGTCACAATTCAGGCTATATATTCCCC
-160
-150
-140
-130
TCATTGGT TCAAATTCGATTCGCTTTCTCATTGGCTGCCC
GCAAAAAAGATAGGTGGTCAACCATTCAAGCCAGCGCACATCGCTT
CCAAAACG CTGCTGGGCATCCATTCAAGTCATCGAACACTGTTA
AGGGCTCA GCCAGTGTCTGTACATACAGCT
3' FLANKING SEQUENCES
(b)
Chicken H2A AGTCCAAA GCTCTTTTCAGAGCCACCCACAGCATCGCAGGAGAGCTC
P.m. H2A AACCTCAACGGCCCTTATCAGGGCCAC CAATT ACTCACGAAAGAAT
P.m. H1 TACCAAAACGGCTCTTTTCAGAGCCAC CAAAT AATCAAGAAAGAAC
P.m. H2B ACATACAACGGCCCTTTTCAGGGCCACACAAAT AATCAAGAAAGAAT

(positions -134, -146, -161) downstream from the putative "TATA box" (Fig. 3a). All show a six-out-of-seven fit to the above consensus sequence. One of these possible cap sites is centred exactly 30 nucleotides from the start of the "TATA box" and for this reason is the most likely site for the initiation of transcription. In the sea urchin histone genes the interval between the "TATA box" and the "cap sequence" varies by up to eight nucleotides[12]. Thus, the other two possible "cap sequences" of the chicken H2A gene at -134 and -161 (Fig. 3a) centred 15 and 42 nucleotides from the "TATA box" cannot be completely excluded as possible 5' mRNA termini. It is possible that all three "cap sequences" are functional and that the H2A transcript has heterogeneous 5' termini.

Assuming that the 5' terminus of the chicken H2A mRNA is in or near the "cap site" at -146 (see Fig. 3a), the mRNA leader sequence (that is, the sequence between the cap site and the AUG) will be 146 nucleotides long and will terminate in the sequence TGCGATG. In *P. miliaris* histone mRNA, the leader sequences terminate in a sequence related to the consensus sequence $^{C}_{A}$CAPyCATG[12]. The chicken H2A gene is clearly an exception to this rule.

The distance from the "cap sequence" to the initiation codon in *P. miliaris* histone genes ranges from 50-70 nucleotides. This is about half

Figure 3a. The 5' flanking regions of chicken histone H2A gene are compared with conserved regions in sea urchin (*P. miliaris*) H2A genes derived from low abundance (h 19) and highly reiterated (h 22) gene clusters respectively. The equivalent region of the chicken ovalbumin gene is also shown[12].

The numbers only refer to the chicken H2A sequence (relative to AUG). The boxed regions show a pentameric sequence, a "TATA box" and the most likely "cap" sequence. Other possible "cap sites" are marked above the sequence in the chicken gene.

Figure 3b. Comparison of 3' untranslated conserved sequences in chicken H2A and three sea urchin (*P. miliaris*) histone genes[14]. Note the reduced homology in the chicken gene of a smaller conserved sequence found in the sea urchin genes.

Numbers in the text refer to the first T of the conserved "TATA box" region,

i.e., 5' $\underset{1}{\text{T}}$ATA 3'

The position of the "cap" sequence is described with reference to the A of the consensus sequence

5' PyCATTCPu 3'.

the size of the probable leader sequence observed for the chicken histone H2A gene. The significance of the increased leader sequence is not known. There are no other initiation codons within the 146 base leader sequence, in accordance with the observations of Kozak[15].

In sea urchin histone genes the "TATA box" is separated by a constant distance (8-10 nucleotides) from a conserved pentameric sequence related to GATCC. The fact that the distance between the "TATA box" and this motif has been conserved supports the proposition that these two elements are stereochemically related[12]. The chicken H2A fragment has the marginally related sequence ACGCC in this position.

Further to the 5' side, Busslinger *et al.*[12] have found a histone gene sequence centred 46 nucleotides upstream from the "TATA box", which is unique to H2A genes of *P. miliaris*. The region of 30 nucleotides contains a dyad symmetry element. In other eukaryotic genes, a region of strong homology, the "CCAAT box" is found in a similar position[16,17]. These regions may be involved in modulation of transcription.

Neither the sea urchin H2A-specific sequence nor the "CCAAT box" are found in the chicken H2A gene. However, in the position where these would be expected, the chicken H2A gene contains a run of 21 A residues intercepted in the centre by one G (42 bases 5' to the "TATA box", Fig. 2). Its particular asymmetry and position relative to other control regions suggest that the A-rich sequence may have relevance to the expression of the H2A gene. The absence of the *P. miliaris* H2A gene specific sequence in the chicken H2A gene contrasts with the remarkable conservation of a 3' non-coding element (see below).

The 3' Flanking Region

Sequencing of the 3' flanking regions of sea urchin histone genes has revealed a 23 base-pair and a 10 base-pair conserved sequence located 29-40 base-pairs from the termination codon[2,14]. These are separated by six base-pairs. By inserting two extra bases between these elements in the chicken gene it is possible to match six out of ten bases of the chicken H2A sequence with the second sea urchin box (Fig. 3b). It appears that strict maintenance of this secondary element is not essential for function.

The chicken H2A gene contains a sequence centred about 45 nucleotides from the termination codon which is remarkably similar to the 23 base-pair sea urchin element (Fig. 3b). This larger homology block contains a region of dyad symmetry. If this sequence were transcribed it could form a stable hairpin loop in the RNA molecule and it has been proposed that this may act as a recognition signal for a regulatory protein or may be involved in maturation or termination of histone DNA transcripts[2,14]. There may be common features of transcription termination for different classes of genes since GC-rich hyphenated dyad symmetry is a feature common to prokaryotic terminator or attenuator sequences[18] as well as eukaryotic polymerase III terminator sequences[19].

Whatever its precise role, the virtually complete conservation of a non-coding region may be significant even at a nucleotide base level rather than simply a secondary structure level since an alternative sequence could give rise to a similar structure. It will be of interest to see whether other vertebrate genes have the same sequence highly conserved, or whether it is histone gene specific.

All 3' termini of sea urchin histone mRNAs were found to map a few nucleotides from the dyad symmetry element. The 3' terminal oligonucleotide of L. pictus H4mRNA has the sequence ACCA-OH[20] and as this sequence can be seen at the end of this highly conserved region it has been proposed that this may form the 3' terminus of histone mRNAs[14].

The sequence AAUAAA is present at high frequency in eukaryotic mRNA trailer sequences[21]. However, this signal is not present in early histone mRNAs from sea urchin and since they are not polyadenylated it has been speculated that it is a polyadenylation signal[3]. This sequence is not found in the 3' trailer region of the chicken H2A gene. This is consistent with results from this laboratory (unpublished) indicating that 90% or more of chicken embryo histone mRNA does not bind to oligo-dT-cellulose.

The presence of conserved sequences in a variety of genes suggests functional roles. While it is difficult to define function by classical genetic analysis in eukaryotes, advances have been made by 'surrogate' genetics in the Xenopus oocyte system[22]. The remarkable conservation of a 3' non-translated region in sea urchin histone genes and the chicken H2A gene suggests a functional role in expression of these genes. Manipulation of this region in conjunction with the oocyte system may provide clues to this role.

ACKNOWLEDGMENTS

We especially thank S. Fields (M.R.C., Cambridge) for subcloning the H2A gene fragment into M13mp7 and for contributing dideoxy sequences. We appreciate help and advice from P. Krieg and A. Robins.

REFERENCES

1. Laine, B., Kmiecik, D., Sautiere, P. and Biserte, G. (1978) Biochemie 60, 147-150.
2. Busslinger, M., Portmann, R. and Birnstiel, M. (1979) Nuc. Acids Res. 6, 2997-3008.
3. Kedes, L.H. (1979) Ann. Rev. Biochem. 48, 837-870.
4. Harvey, R.P. and Wells, J.R.E. (1979) Nuc. Acids Res. 7, 1787-1797.
5. Harvey, R.P., Krieg, P.A., Coles, L. and Wells, J.R.E. (1981) (submitted for publication).
6. Maxam, A. and Gilbert, W. (1979) Methods in Enzymology, Grossman, L. and Moldave, K., eds. Vol. 65, part 1, pp. 449-560, Academic Press, N.Y.
7. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Nat. Acad. Sci. USA 74, 5463-5467.
8. Anderson, S., Gait, M.J., Mayol, L. and Young, I.G. (1980) Nuc. Acids Res. 8, 1731-1743.
9. Sinclair, J.H. and Brown, D.D. (1971) Biochemistry 10, 2761-2769.
10. Richards, R.I., Shine, J., Ullrich, A., Wells, J.R.E. and Goodman, H.M. (1979) Nuc. Acids Res. 7, 1137-1146.
11. Richards, R.I. and Wells, J.R.E. (1980) J. Biol. Chem. 255, 9306-9311.
12. Busslinger, M., Portmann, R., Irminger, J.C. and Birnstiel, M.L. (1980) Nuc. Acids Res. 8, 957-977.
13. Corden, J., Wasylyk, B., Buchwalcher, A., Sassone-Corsi, P., Kedinger, C. and Chambon, P. (1980) Science 209, 1406-1414.
14. Hentschel, C., Irminger, J., Bucher, P. and Birnstiel, M.L. (1980) Nature 285, 147-151.
15. Kozak, M. (1978) Cell 15, 1109-1123.
16. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) Nuc. Acids Res. 8, 127-142.
17. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) Cell 21, 653-668.
18. Adhya, S. and Gottesman, M. (1978) Ann. Rev. Biochem. 47, 967-997.
19. Korn, L.J. and Brown, D.D. (1978) Cell 15, 1145-1156.
20. Grunstein, M. and Schedl, P. (1976) J. Mol. Biol. 104, 323-349.
21. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
22. Grosschedl, R. and Birnstiel, M.L. (1980) Proc. Nat. Acad. Sci. USA 77, 1432-1436.