Volume 11 Number 19 1983 Nucleic Acids Research

Structure of a cluster of mouse histone genes

Donald B.Sittman*, Reed A.Graves and William F.Marzluff

Department of Chemistry, Florida State University, Tallahassee, FL 32306, USA

Received 21 July 1983; Revised and Accepted 13 September 1983

ABSTRACT

The four mouse histone genes(2 H3 genes, an H2b gene and an H2a gene) present in a cloned 12.9 kilobase fragment of DNA have been completely sequenced including both 5' and 3' flanking regions. These genes are expressed in cultured mouse cells and the 3' and 5' ends of the mRNA have been determined by S1 nuclease mapping. These genes code for a minor fraction of the histone mRNAs expressed in cultured mouse cells. They comprise at most 5-8% of the total histone mRNA of each type. The two H3 genes code for H3.2 and H3.1 histone proteins, while the H2b gene codes for an H2b.1 protein with a single amino acid change (val-leu)at position 18. Only the 3' portion of the H2a gene is contained in the clone and there is an amino acid change (alanine - proline) at position 126. Comparison of the 5' and 3' flanking sequences reveals a conserved sequence at the 3' end of the mRNA which forms a hairpin loop structure. The codon usage in the genes is non-random and there has been no discrimination against CG doublets in the coding region of the genes.

INTRODUCTION

Histone genes have been isolated from a broad range of species.(1-7) In those species where there is an early development characterized by a rapid cleavage stage e.g. sea urchin and drosophila, there is a high gene copy number.(1,2) A majority of the histone genes in these organisms code for histones which are expressed early in development(1). It has been suggested that these genes are in high copy number to provide sufficient histones for the rapid cell divisions of early embryogenesis (1). The histone genes of higher eukaryotes (mammals and birds) and primitive eucaryotes (yeast) are present in lower copy numbers (4-6). The genes in higher eucaryotes appear to be arranged as clusters with no apparent order to the position of particular histone genes within these clusters (4-6). Recently the genes

© IRL Press Limited, Oxford, England.

coding for the "late" histones of sea urchin have been isolated (8,9) and they also appear to be arranged as random clusters.

Histones, although considered to be highly conserved during evolution (10) are nevertheless comprised of a fairly large diversity of variant forms (10), with slight differences in amino acid sequence. Zweidler and coworkers (11) have characterized three groups of mammalian histones: 1. Replication variants (H2a.1, H2b.2, H3.1, H3.2, H1a and H1b) which predominate in rapidly dividing cells. 2. Replacement variants (H2a.2, H2a.3, H2b.1, H3.3 and H1-0) which are found in larger amounts in differentiated cells as well as in variable amounts between different cell types. 3. Spermatocyte variants which predominate during spermatogenesis.

We previously reported the isolation of two clusters of mouse histone genes (5). We report here the sequences and mRNA structures of four genes on one of these clones, MM221.

METHODS

DNA cloning and sequencing

The following DNA subclones were prepared from the histone gene cluster MM221(5): pMH3.2, a 0.9kb Eco R1-Sal I fragment was cloned into pBR322 from which the 650 bp Eco R1-Sal I fragment had been removed; pMH3.1, a 2.4kb Eco R1-Sal I fragment inserted into the same vector; pMH2b-H3, a 6.0kb Eco R1 fragment inserted into the Eco R1 site of pACYC 184(12); pMH2a, a 1.1kb Eco R1 fragment inserted into pACYC 184; pMH2b, a 1.2kb Sma I-Hinc II fragment inserted into the Sma I site of pUC8(13).

The DNA fragments were sequenced by the method of Maxam and Gilbert(14,15). Routinely six sequencing reactions were used: a G reaction, a G + A reaction, an A>C reaction, a C + T reaction and a C reaction described by Maxam and Gilbert(15) and the T specific reaction described by Rubin and Schmid(16) which in our hands also reacts with G.(17)

The DNA was end-labelled with polynucleotide kinase at the 5' end or with the Klenow fragment of DNA polymerase I at the 3' end. The DNA was then cleaved with a second restriction enzyme and the singly end-labeled fragment prepared by polyacrylamide gel electrophoresis. After elution from the gel many of the

fragments were purified by binding them to a small column of NACS(Bethesda Research Labs). The fragments were eluted from the resin, and precipitated with ethanol prior to sequencing. All sequences were done at least twice and greater than 90% of the sequence was confirmed by sequencing both strands of the DNA.

The sequences were analyzed using the computer programs of the Molgen group on the Sumex system at Stanford University.

S1 Nuclease Digestion

S1 nuclease was from P/L biochemicals and required no further purification. DNA probes were as indicated in the figure legends. Ten micrograms of cytoplasmic RNA were precipitated with the indicated DNA probes and dried. Nucleic acids were then dissolved in 5 microliters of 80% formamide buffer(18). RNA-DNA hybrids were then formed by denaturation at 85°C for 15 minutes followed by a 3 hour incubation at 58°C. Ninety-five microliters of ice-cold S1 buffer (19) was added followed by S1 nuclease to a final concentration of either 2,000 units/mL or 200 units/ml. Nuclease digestion was for one hour at room temperature except when H2A DNA was used as a probe, in which case digestion was at 12°C for 1 hour. After digestion the protected DNA was recovered by ethanol precipitation and analyzed by polyacrylamide gel electrophoresis in 8.2M urea, as previously described.(20)

MATERIALS

Restriction enzymes and enzymes used in DNA sequencing were obtained from Bethesda Research Laboratories. Radiochemicals were purchased from ICN(γ-$^{32}PO_4$ATP) or from Amersham α-$^{32}PO_4$dCTP).

RESULTS

Sequence analysis.

Figure 1 shows a restriction map of MM221 and the subcloned fragments from this clone which were used for sequencing. pMH3-2 is a 0.9 kb Eco R1, Sal I restriction fragment cloned into pBR322 containing the 3' end of the H3.2 gene starting at the Sal I site at codon 57. The 5' end of the H3.2 gene and the complete H2b gene are contained in the plasmid pMH2b, a 4.9kb EcoR1, Sal I fragment cloned into pBR322. The complete H3.1 gene is contained

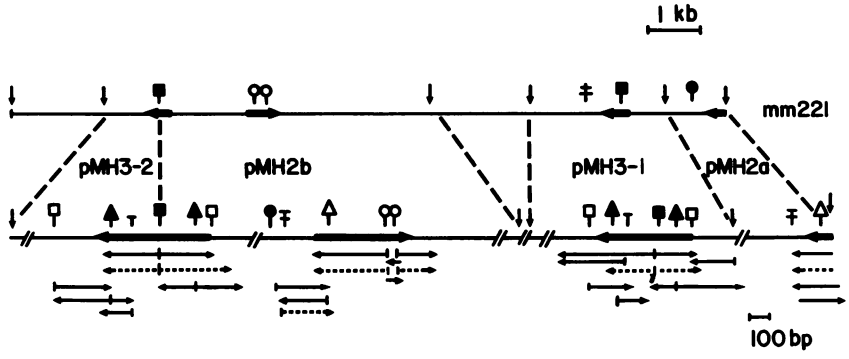


Figure 1. Map of the MM221 Phage.
The structure of the 12.9 kb insert in MM 221 and the subclones derived from it is shown. The location and orientation of the histone genes is indicated. The sequencing strategies used are indicated by the arrows below the map. The dotted lines indicate fragments sequenced by 3' labeling and solid lines indicate fragments sequenced by 5' labeling. The restriction enzymes sites are indicated: ↓Eco Rl; Sal I; Ava I; Ava II; Dde I; Hinf I; Sau 96 I; Pvu II; Pst I. Only those sites used in the sequencing strategies are indicated.

in the plasmid pMH3-1 which is a 2.4 kb EcoRl fragment inserted into pACYC184 (11). pMH2a contains a 1.1 kb EcoRl fragment containing the 3' end of an H2a gene starting at codon 93 inserted into pACYC184. The EcoRl site within this gene was generated by insertion of a linker into an Alu I site during the construction of the gene library (21).

The sequencing strategy used is indicated by arrows in Figure 1. We used both 3' and 5' end labeling and 95% of the sequences have been confirmed by sequencing both strands. The rest of the sequences have been done at least twice.

The sequences of the coding and flanking regions of the H3.1 and H3.2 genes are presented in Figure 2. The sequences of the coding regions of these two genes are very similar. There are only 13 base differences within the coding region of these two genes making them 97% homologous. All of the base changes are silent except one; this one accounts for the cysteine to serine change at amino acid 96. The regions immediately flanking the coding region of the H3.1 and H3.2 genes have little detectable homology. The initiation and termination sites of the mRNA have

H3 GENES

```
      180                                 150
3.2                                     C ACAAATTTGA AGTTGAGACC TGTTATCCAA
3.1   CTCAGCCAAT AGGACTACTG CGCGGGACAC TTGAAAAGCA GACACGCCTA TCAGGATGCT

      120                                 90
3.2   TTACCAAGTA CTTCCGCATA CATCATCATA GGCATTTGAA GATTTCAACC AATCAGGAGC
3.1   TTCTCGGTGG GAAGGAGGGG TACGAGCGCG GGTACGTGTG TTGCGCGTGT GCGACGCAAG
      60                                  30       *
3.2   ATGTTCCTTC TATAAAGGAA CCCAGAACCT AACCTCTGCA TTTCCTATTT CTTTGTAGAA
3.1   CGTACTTAAA GGCCAAAGTG CGCTACTTAG GTATCTCACT TTTCCCTACG GTTACTTGCC
                                            *

H3-2                                                C   T
H3-1  ATG GCT CGT ACT AAG CAG ACC GCT CGC AAG TCT ACC GGC GGC AAG GCC CCG CGC
      MET Ala Arg Thr Lys Gln Thr Ala Arg Lys Ser Thr Gly Gly Lys Ala Pro Arg

H3-2
H3-1  AAG CAG CTG GCC ACC AAG GCC GCC CGC AAG AGC GCC CCG GCC ACC GGC GGC GTG
      Lys Gln Leu Ala Thr Lys Ala Ala Arg Lys Ser Ala Pro Ala Thr Gly Gly Val

H3-2        A                                   C       G
H3-1  AAG AAG CCT CAC CGC TAC CGT CCC GGC ACT GTG GCA CTG CGC GAG ATC CGG CGC
      Lys Lys Pro His Arg Tyr Arg Pro Gly Thr Val Ala Leu Arg Glu Ile Arg Arg

H3-2
H3-1  TAC CAG AAG TCG ACC GAG CTG CTG ATC CGC AAG CTG CCG TTC CAG CGC TTG GTG
      Tyr Gln Lys Ser Thr Glu Leu Leu Ile Arg Lys Leu Pro Phe Gln Arg Leu Val

H3-2
H3-1  CGC GAG ATC GCG CAG GAC TTC AAG ACC GAC CTG CGC TTC CAG AGC TCG GCC GTC
      Arg Glu Ile Ala Gln Asp Phe Lys Thr Asp Leu Arg Phe Gln Ser Ser Ala Val

H3-2                          Ser
H3-2                        G A C                 T
H3-1  ATG GCT CTG CAG GAG GCC TGT GAG GCC TAC CTC GTG GGT CTG TTT GAG GAC ACC
H3-1  Met Ala Leu Gln Glu Ala Cys Glu Ala Tyr Leu Val Gly Leu Phe Glu Asp Thr

H3-2
H3-1  AAC CTG TGC GCC ATC CAC GCC AAG CGT GTC ACC ATC ATG CCC AAG GAC ATC CAG
      Asn Leu Cys Ala Ile His Ala Lys Arg Val Thr Ile Met Pro Lys Asp Ile Gln
                                                                          20
H3-2            C   T           C     C             TAGGCACGCT TTCTACACTG
H3-1  CTG GCC CGT CGC ATC CGC GGG GAG AGG GCT TAA  GGGTTTCTGT TAATCCACAC
      Leu Ala Arg Arg Ile Arg Gly Glu Arg Ala  .
                                      50   ***                          80
H3-2  GCACGTAAAC CAAAACGGCT CTTTTAAGAG CCACCTCCAT TATCCACCAA AGATGCTTGA
H3-1  AACCACTTTA AAGGCTCTTC TTAGAGCCAC CCATCTTCCA AAAAAAGAAC TGTGCGCTTT
                                     ** *
                                     110                              140
H3-2  AGTACAAGTT GTGAGAGTTT TCTAGGGTTT CCTATTATAG CCTTTCTTGA CAATGTGAGC
H3-1  TTCCAAACTT GTGGGTATTA ATCAGTTTCA TTTGTCAAAA GTGCTAGGTC TCC
```

Figure 2. Sequence of the 2 H3 genes.
The sequence of the H3.1 and H3.2 genes are indicated. Only the bases which differ between H3.2 and H3.1 genes are shown in the coding region. The beginning and end of the mRNA as determined by S1 nuclease mapping are marked with an asterisk.

```
H2B GENE

                                    150
            TGATACTTAC GCAACTAACC ACTGAGCGAA TATGCTTCCT TGATGGACAG

120                                 90
TTAGTGCTTG ACGTTTGCAG ACTCTCTGAC AAGGACAGCC ACCGCTTTAT TTAAAGAGCA

60                    **            30
GGAAAGGAAC GGAACAGTTC AATATCTCTT TTCTTGGCCT ACCTTCATTC TCTGTTCACT

ATG CCT GAG CCC GCC AAG TCC GCT CCT GCC CCG AAG AAG GGC TCC AAG AAG GCC
MET Pro Glu Pro Ala Lys Ser Ala Pro Ala Pro Lys Lys Gly Ser Lys Lys Ala

CTG ACC AAG GCC CAG AAG AAG GAC GGC AAG AAG CGC AAG CGC AGC CGC AAG GAG
Leu Thr Lys Ala Gln Lys Lys Asp Gly Lys Lys Arg Lys Arg Ser Arg Lys Glu

AGC TAC TCG GTG TAC GTG TAC AAG GTG CTG AAG CAA GTG CAC CCC GAC ACC GGC
Ser Tyr Ser Val Tyr Val Tyr Lys Val Leu Lys Gln Val His Pro Asp Thr Gly

ATC TCC TCC AAG GCC ATG GGC ATC ATG AAC TCG TTC GTG AAC GAC ATC TTC GAG
Ile Ser Ser Lys Ala Met Gly Ile Met Asn Ser Phe Val Asn Asp Ile Phe Glu

CGC ATC GCG GGA GAG GCG TCC CGC CTA GCG CAT TAC AAC AAG CGC TCG ACC ATC
Arg Ile Ala Gly Glu Ala Ser Arg Leu Ala His Tyr Asn Lys Arg Ser Thr Ile

ACG TCC CGG GAG ATC CAG ACG GCC GTG CGC CTG CTG CTG CCC GGG GAG CTG GCC
Thr Ser Arg Glu Ile Gln Thr Ala Val Arg Leu Leu Leu Pro Gly Glu Leu Ala

AAG CAC GCC GTG TCG GAG GGC ACC AAG GCT GTC ACC AAG TAC ACC AGC TCC AAG
Lys His Ala Val Ser Glu Gly Thr Lys Ala Val Thr Lys Tyr Thr Ser Ser Lys
                          20                                    ***50
TGA GTGCTCAAGA  CTCAGCTCTT  AACCCAAAGG  CTCTTTTCAG  AGCCACTCAA

                                    80                                  110
GACTTCAAAA  TTGGAGCTTT  AATGCTACCA  AGCGACTTAG  TGACTACCGG  GAAAATAACC

                                  137
GACTTCATCG  CAGGATGTGT  ACAACAC
```

Figure 3. H2b Sequence.
The sequence of the H2b gene is shown. The beginning and end of the mRNA is indicated with an asterisk.

been determined by S1 nuclease mapping (see below) and are indicated in Figure 2 with an asterisk.

The sequence of the mouse H2b gene from MM221 codes for a protein that differs from the calf H2b by one amino acid (Fig. 3). The calf has a valine at position 18(10) while the H2b from MM221 has a leucine at this position. Franklin and Zweidler have reported three variant forms of mammalian H2b none of which have a change reported at this position(11). It is likely that the valine to leucine change would not be detected by gel electrophoresis, particularly since this gene codes for only a small fraction of the total H2b RNA. The sequence of the gene indicates that it codes for an H2b.1 variant.

Only the 3' end of the H2a gene was obtained in the clone MM221. The MM221 clone was isolated from a library which was generated by partial Alu I and Hae III digestion of total genomic

```
H2A GENE
92                                100
CTG AAC AAG CTG TTG GGC CGC GTG ACC ATC GCG CAG GGC GGC GTC CTG CCC AAC
Leu Asn Lys Leu Leu Gly Arg Val Thr Ile Ala Gln Gly Gly Val Leu Pro Asn

110                                   120
ATC CAG GCC GTG CTG CTG CCC AAG AAG ACC GAG AGC CAC CAT AAG CCC AAG GGA
Ile Gln Ala Val Leu Leu Pro Lys Lys Thr Glu Ser His His Lys Pro Lys Gly
128                             20                                    50
AAG TAA    GCCAGTGAGC TAAGTTTTTT TTTTTTTTTT TTTTTTTTTT TAAACAAAAC
Lys  .

                           *    80
CCAAGGCTCT TTTCAGAGCC ACCACTTCTT CATATAAGAG
```

Figure 4. H2a Sequence.
The sequence of the portion of the H2a gene present in MM 221 is shown. The end of the mRNA determined by S1 nuclease mapping is shown.

DNA followed by blunt end ligation of Eco R1 linkers to the cleaved sites prior to insertion into charon 4A (21). Sequencing indicated that an Alu I site was cleaved at codon 92 in the formation of this recombinant phage. We therefore obtained only the region coding for the terminal 37 amino acids of this gene (Fig. 4). As with the mouse H2b gene derived from MM221 this H2a gene appears to code for a protein variant different from the three H2a variants described by Franklin and Zweidler.(10) The H2a from MM221 codes for a histone which most closely resembles an H2a.1 variant. It differs from other H2a proteins in having a proline at position 126 instead of an alanine.

S1 Nuclease Mapping.

We have mapped the start and end of the mRNA derived from the histone genes in MM221 by S1 nuclease mapping (19,22). Figure 5 presents a low resolution map of the size of 5' labeled H3.1, H3.2 and H2b fragments protected from S-1 digestion by polysomal RNA. Identical results were obtained with total cell RNA. For this analysis the genes were 5' labelled at the internal Sal I site (amino acid 57) of the H3 genes and at the Ava I site (amino acid 92) of the H2b gene. The major protected bands for the H3.1 and H3.2 genes (figure 5, lanes 1 and 2 respectively) are the size expected, 175 bases from the Sal I site, for protection by the mRNA to up to the AUG codon. The slower migrating bands seen above the 175 bp band represent the protection of

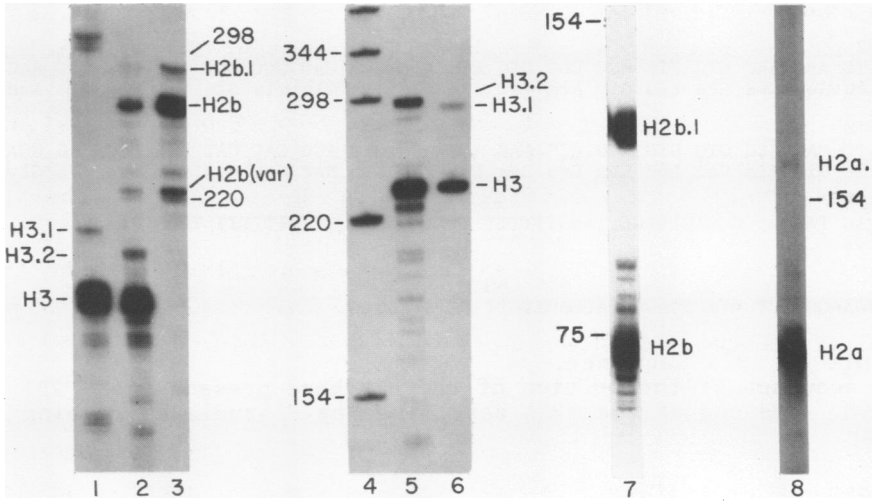


Figure 5. S1 Nuclease Mapping of the Histone Genes.
The various histone genes were end-labeled at an internal restriction enzyme site and hybridiidzed to cytoplasmic RNA. The hybrids were treated with S1 nuclease and the protected DNA analyzed by polyacrylamide gel electrophoresis as described in Materials and Methods. pBR 322 digested with Hinf was end-labeled and used as a size marker. Lane 1: H3.1 labeled at the 5' end of the Sal I site (amino acid 57). The protected bands are H3, protection by a family of mRNAs similar in sequence in the coding region and H3.1 protection by the mRNA derived from this gene. Lane 2: H3.2 labeled at the 5' end of Sal I site at amino acid 57. The same H3 band is protected as in lane 1. The H3.2 band is the DNA protected by the mRNA derived from this gene. Minor bands of large molecular weight are due to contamination with the end labeled H2b.1 fragment. Lane 3: The H2b.1 gene was end-labeled at the 5' end of the Ava I site at amino acid 91-92. The protected bands are H2b(var)- protection to amino acid 18 where there is an amino acid change; H2b-protection by a family of mRNAs which are similar in the coding region up to the AUG codon; H2b.1-protected by the mRNA derived from this gene. Lane 4: pBR 322 marker. Lane 5: The H3.2 gene was labeled at the 3' end of the Sal I site. The protected bands are H3-protection by a family of mRNA with similar coding regions to the TAA codon; H3.2-protection by the mRNA derived from this gene. Lane 6: The H3.1 gene was labeled at the 3' end of the Sal I site. The protected bands are H3 as in Lane 5 and H3.1-protection by the mRNA derived from this gene. Lane 7: The H2b gene was labeled at the 3' end of the Ava I site at amino acid 104. The protected bands are H2b- protection by a family of H2b mRNAs up to the TAA codon and H2b.1 protection by the mRNA derived from this gene. Lane 8: The H2a gene was labeled at the 3' end of the Sau 96 I site at amino acid 97. The protected bands are H2a-protection by a family of mRNAs similar in the coding region up to the TAA codon and H2a.1 protection by the mRNA derived from this gene.

the DNA by the mRNAs transcribed from these genes. The 5' sequences flanking the coding region share very little homology between the H3.1 and H3.2 genes (Fig. 2A), so only the mRNA derived from the H3.2 gene will protect the H3.2 DNA for a significant distance past the AUG codon. This assumes there are no other H3 genes with a sufficiently high homology to either of these genes to protect transcripts beyond the AUG codon. The H2b probe protects three DNA fragments. The smallest maps to the amino acid change at position 18.(unpublished results) The major one maps to the AUG codon and the largest fragment maps to the start of the mRNA derived from this gene.

We also mapped the termination sites of the mRNAs derived from the four genes present in MM221. The 3' ends were mapped in an analogous manner to the 5' ends except that the internal sites were 3' labelled by filling in the ends with the Klenow fragment of DNA polymerase I. Figure 5 shows the results. The major bands protected in all 4 genes maps to the termination codon, similar to the result with the 5' end mapping. The H3.1 and H3.2 genes terminate at approximately an equal distance, 50 bases from the translation termination signal, the H3.2 mRNA extending slightly further. The termination site for the H2b gene is about 45 bases from the translation termination signal. The termination site for the H2a transcript was determined by 3' labeling the internal Sau 961 site at amino acid 98 in the H2a gene. The H2a transcript protecting the gene from MM221 terminates approximately 75 bases from the translation termination signal after the 27 consecutive thymidine residues in the 3' region of the DNA. This fragment was only detected when S1 nuclease digestion was done at 10 degrees presumably because the AT hybrid is not stable to S1 nuclease digestion at higher temperatures. These genes each code for only a small fraction of the histone mRNAs in mouse myeloma cells. We estimated the proportion of the mRNA derived from these genes by densitometry of the autoradiograms, comparing the intensity of the band at the AUG or TAA codon with the band at the terminus of the mRNA. Each of these genes codes for no more than 5% of the mRNA for that particular class of histone.

To determine the precise termini of the mRNAs, high resolu-

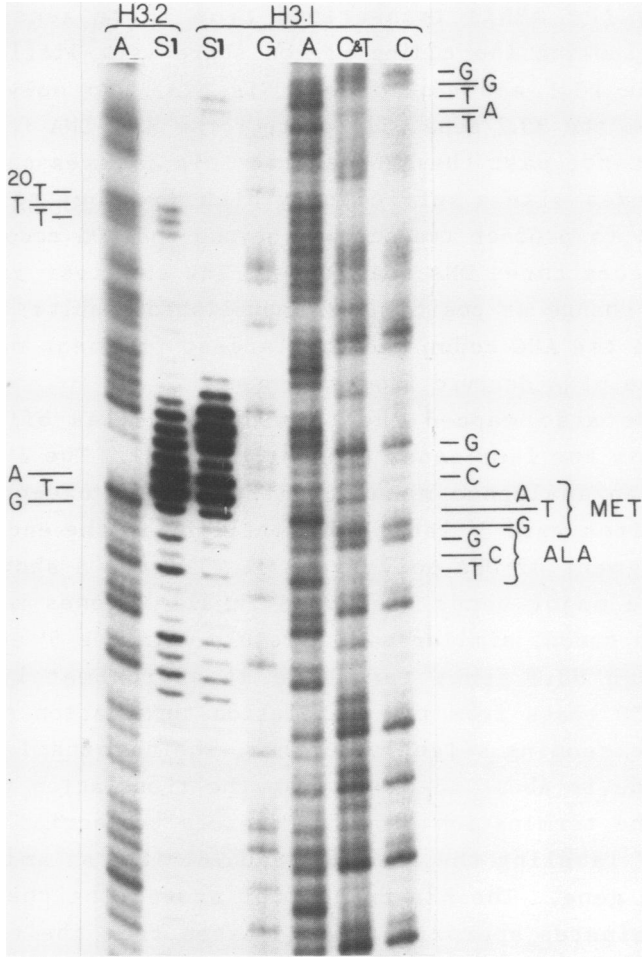


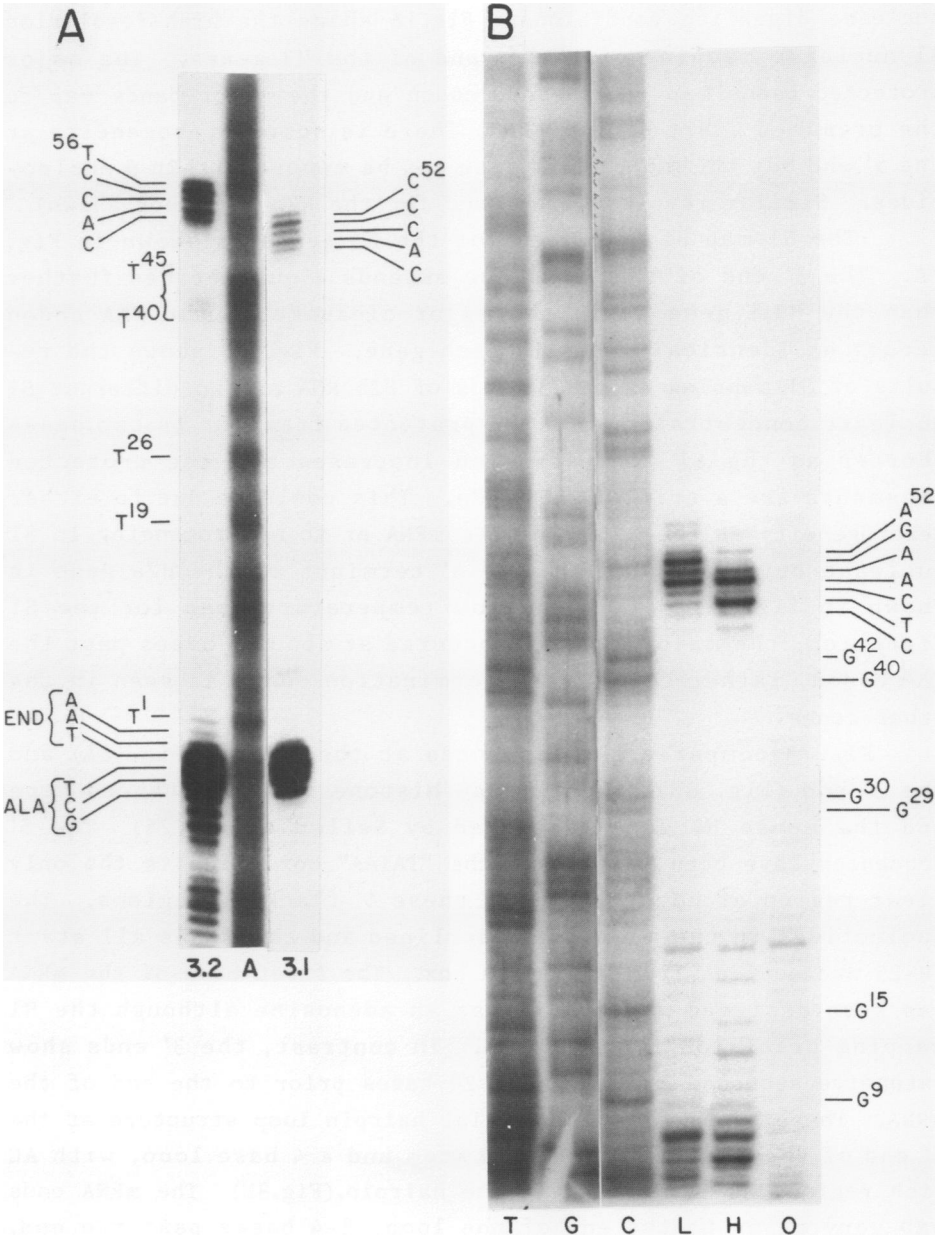
Figure 6. 5' Ends of the H3 Genes.
The 5' ends of the H3.1 and H3.2 genes were mapped by 5' labeling the Sal I site at amino acid 57. The S1 nuclease assays were analyzed next to DNA sequencing reactions. Lane 1 is an A reaction of H3.2 DNA; The position of T's in the complementary strand is indicated. The numbers refer to the distance from the AUG codon. Lane 2 - S1 assay - H3.2; Lane 3 - S1 assay - H3.1; Lane 4 -7 Sequencing reactions of the H3.1 DNA. Left to right: G, A, C+T, C.

tion S1 mapping has been carried out on all four genes. The resolution of this method is several bases, due both to the multiplet of bands observed and the specificity of S1 nuclease.(23,24) The results also vary with changes in the S1

nuclease digestion conditions. Fig. 6 shows the high-resolution S1 nuclease mapping of the 5' end of the H3 genes. The major protected bands map to the AUG codon and the minor bands map to the presumed start of the mRNA. There is some heterogeneity at the 5' end but the mRNA start site can be mapped within 3 nucleotides. Similar results were found for the H2b gene.(not shown)

The S1 map of the 3' end of the H3 genes is shown in Fig. 7A. The 3' end of the H3.2 gene extends 4 nucleotides further than the H3.1 gene while the major cleavage at the TAA codon occurs at identical sites in each gene. Fig. 7B shows the results of S1 mapping of the 3' ends of H2b RNA at two different S1 nuclease concentrations. The protected fragment is two bases shorter as the S1 concentration increases and the protected fragments are a triplet of bands. This could be due to either heterogeneity at the 3' end of the mRNA or to heterogeneity in S1 nuclease cutting or both. The 3' terminus of the H2a gene is shown in figure 7C. At the low temperature used for the S1 digestion, the major cleavage occured at about 5 bases past the TAA codon, rather than at the termination codon as seen in the other genes.

Fig. 8 compares the sequences at the 5' end (Fig. 8A) and the 3' end (Fig. 8B) of the mouse histone genes sequenced here and the mouse H4 gene sequenced by Seiler et al.(25) The 5' sequences have been aligned at the "TATAA" box which is the only clear region of homology among these 5' flanking regions. The nucleotides in the mRNA are underlined and the mRNAs all start 20-25 nucleotides from the TATAA box. The first base of the mRNA has been assigned arbitrarily as an adenosine although the S1 mapping precision is 3 bases. In contrast, the 3' ends show extensive sequence homology for 20 bases prior to the end of the mRNA. These bases form a potential hairpin loop structure at the 3' end of the mRNA with a 6-base stem and a 4 base loop, with AC rich regions at either end of the hairpin.(Fig.8B) The mRNA ends map very close to the end of the loop, 2-4 bases past the end. The sequences are identical in the hairpin, except in the loop, in all 4 genes from MM221. The H4 gene whose location in the genome relative to MM221 is not known, has two bases changed in the stem which changes an AT pair to a GC pair preserving the


A
56
T
C
C
A
C
T45
T40
T26
T19
END
A
A
T
T1
ALA
T
C
G
C52
C
C
A
C
3.2
A
3.1
B
A52
G
A
A
C
T
C
G42
G40
G30
G29
G15
G9
T
G
C
L
H
O

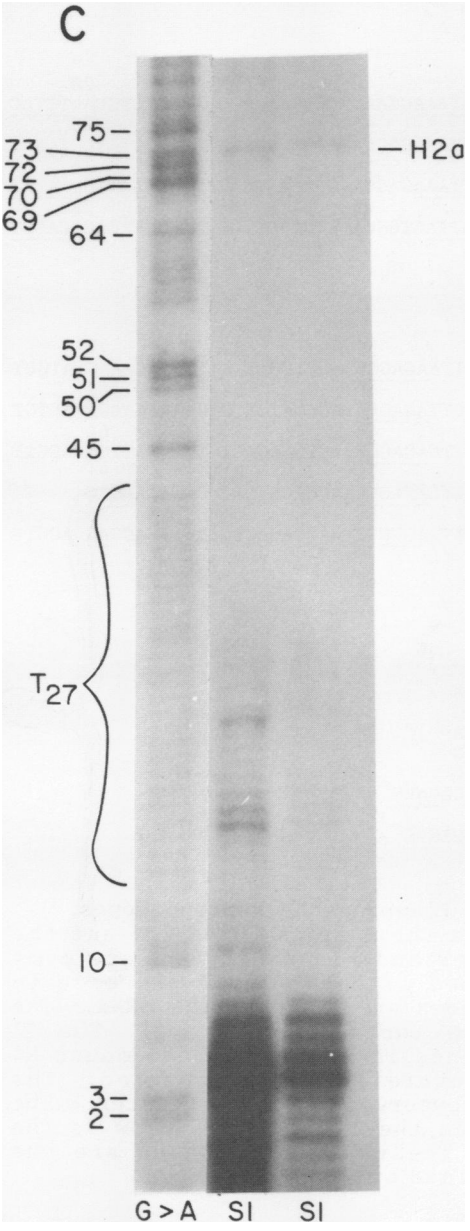


Figure 7. 3' End of the Histone Genes.
A. The 3' ends of the H3.2 and H3.1 genes were mapped by 3' labeling at the Sal I site at amino acid 57. The S1-resistant DNA was analyzed on a sequencing gel. The center lane shows an A specific reaction of the H3.2 DNA. The sequence around the S1 sites is indicated. B. The H2b gene was labeled at the 3' of the Ava I site of amino acid 104. The T,G and C lanes are DNA sequencing reactions. The numbers refer to the distance of the G's in the complementary strand from the TAA codon. The S1 lanes show the protected DNA after treated with low(L) and high(H) concentrations of S1 nuclease respectively. Lane O is an S1 nuclease reaction done with tRNA. C. The H2a gene was labeled the 3' end of the Sau 96 I site at amino acid 97. On the right is the G reaction. The position of the C's in the complementary strands is indicated. The numbers are the distance from the TAA codon. The next two lanes show two different S1 nuclease assays.

A

5' FLANKING HOMOLOGIES

```
            70          60          50          40          30          20
H3-2  ATTTCAACCA ATCAGGAGCA TGTTCCTTCT ATAAAGGAAC CCAGAACCTC AACCTCTGCA TTTCC
            75          65          55          45          35          25
H3-1  GTGTGTTGCG CGTGTGCGAC GCAAGCGTAC TTAAAGGCCA AAGTGCGCTA CTTAGGTATC TCACT
            48          58          68          58          48          38
H2B   ACTCTCTGAC AAGGACAGCC ACCGCTTTAT TTAAAGAGCA GGAAAGGAAC GGAACAGTTC AATAT
            78          68          58          48          38          28
H4    AGTTTTCAAT CTGGTCCGAT CCTCTCATAT ATTAGTGGCA CTCCACCTCC AATGCCTCAC CAGCT
```

B

3' FLANKING HOMOLOGIES

```
            22          32          42          52          62          72
H3-2  CTACACTGGC ACGTAAACCA AAACGGCTCT TTTAAGAGCC ACCTCCATTA TCCACCAAAG ATGCT
            18          28          38          48          58          68
H3-1  GTTAATCCAC ACAACCACTT TAAAGGCTCT TCTTAGAGCC ACCCATCTTC CAAAAAAAGA ACTGT
            14          24          34          44          54          64
H2-B  TCAAGACTCA GCTCTTAACC CAAAGGCTCT TTTCAGAGCC ACTCAAGACT TCAAAATTGG AGCTT
            41          51          61          71          81          91
H2-A  TTTTTTTTTT TAAACAAAAC CCAAGGCTCT TTTCAGAGCC ACCACTTCTT CATATAAGAG
            26          36          46          56          66          76
H4    TCCCCCCCCC CCCCCATCCC TAACGGCCCT TTTTAGGGCC AACCACAGTC TCTTCAGGAG AGCTG
```

```
   T - T                    N - T
  /     \                  /     \
 T       T                T       N
  \     /                  \     /
   T-A                      T-A
   C-G                      C-G
   C-G                      T-A
   C-G                      C-G
   G-C                      G-C
TAACG-CAACC              CAAAG-CACNNN

H4 GENE                  MM221 GENES
```

Figure 8. Comparison of Sequences Flanking the Histone Genes. A. The 5' flanking sequences from the 3 genes in MM 221 and the mouse H4 gene sequenced by Seiler and Birnsteil (25) are compared. The ends have been aligned with the TATAA sequence in register. The underlined nucleotides are found in the mRNA. The numbers refer to distances from the AUG codon. B. The 3' flanking sequences from the 4 genes in MM 221 and the mouse H4 gene sequenced by Seiler and Birnstiel (25) are compared. The ends have been aligned with the conserved dyad in register. The numbers refer to the distances from the TAA codon. Below is the dyad structure for the genes. The underlined bases are the likely end of the mRNAs as determined by S1 mapping.

stem structure. Similar structures are found at the 3' end of other histone genes(26,27).

A striking feature of the histone coding regions is their high GC content. This is due primarily to the non-random codon

TABLE IA
DINUCLEOTIDE FREQUENCY IN HISTONE CODING REGIONS

	CODING REGION (64.7% GC)			
	Number	Percent	Expected	Percent
AA	85	5.2	4.4	
AC	84	5.1	7.1	
AG	130	8.0	6.5	
AT	43	2.6	3.0	
CA	131	8.0	7.1	
CC	169	10.4	11.5	
CG	157	9.6	10.4	
CT	96	5.9	4.9	
GA	95	5.8	6.5	
GC	209	12.8	10.4	
GG	122	7.5	9.5	
GT	75	4.6	4.5	
TA	31	1.9	3.0	
TC	89	5.5	4.9	
TG	94	5.8	4.5	
TT	22	1.3	2.1	

The dinucleotide frequencies for the coding regions(starting at the AUG and the termination codon) are given along with the expected values based on the base composition. The genes analyzed were the genes in MM221 and the H4 gene sequenced by Seiler et al.(25).

TABLE IB
CODON USAGE IN MOUSE HISTONE GENES

TTT-Phe 2	TCT-Ser 2	TAT-Tyr 0	TGT-Cys 1
TTC-Phe 10	TCC-Ser 9	TAC-Tyr 15	TGC-Cys 2
TTA-Leu 0	TCA-Ser 0	TAA- . 3	TGA- . 1
TTG-Leu 4	TCG-Ser 8	TAG- . 1	TGG-Trp 0
CTT-Leu 1	CCT-Pro 4	CAT-His 4	CGT-Arg 10
CTC-Leu 4	CCC-Pro 11	CAC-His 8	CGC-Arg 42
CTA-Leu 2	CCA-Pro 0	CAA-Gln 1	CGA-Arg 0
CTG-Leu 34	CCG-Pro 7	CAG-Gln 22	CGG-Arg 5
ATT-Ile 0	ACT-Thr 4	AAT-Asn 0	AGT-Ser 0
ATC-Ile 28	ACC-Thr 31	AAC-Asn 9	AGC-Ser 9
ATA-Ile 0	ACA-Thr 0	AAA-Lys 2	AGA-Arg 1
ATG-MET 11	ACG-Thr 2	AAG-Lys 61	AGG-Arg 1
GTT-Val 0	GCT-Ala 13	GAT-Asp 1	GGT-Gly 8
GTC-Val 11	GCC-Ala 35	GAC-Asp 13	GGC-Gly 26
GTA-Val 0	GCA-Ala 2	GAA-Glu 0	GGA-Gly 6
GTG-Val 21	GCG-Ala 9	GAG-Glu 25	GGG-Gly 3

The codon frequencies for the genes in MM221 and the mouse H4 gene are given.

usage (Table IA). The 3rd bases are predominently C or G. In addition there has been no selection against CG doublets, which occur in the predicted frequency (Table IB) in the coding region. This is due partly to the almost exclusive use of the arginine codons CGX in these relatively arginine-rich proteins as well as to many CG doublets both in the last two bases of alanine, serine and proline codons as well as between the 3rd and 1st bases of adjacent codons. This situation is in marked contrast to the codon usage in the human globin mRNAs (28) and serum albumin mRNA(29) where there are few CG doublets in the codons used. The preferred arginine codons in these genes are the AGQ codons. In the codons NCX, the X base is rarely G although G is used frequently as the third base in other codons.(28,29) Analysis of the codon usage in the chicken histone H2a and H2b sequences which have been reported(30,31) reveals the same pattern of codon preference with no discrimination against codons which contain CG doublets. In contrast, analysis of the early sea urchin(S. purpuratus) H2a, H2b and H3 genes(32) reveals very few codons containing CG doublets. This suggests that the evolutionary forces which act to reduce the frequency of CG doublets in mammalian genomes do not affect the coding regions of the mouse histone genes. The unusual codon usage in the histone genes from both mouse and chicken suggests that this selectivity may play a functional role in histone mRNA or histone gene structure.

Thus the histone genes of the mouse studied here have several features in common: 1. there are no intervening sequences in the coding region, 2. there is a common 3' structure at the end of the mRNA similar to that in sea urchin(26), chicken (27), drosophila(26) and xenopus histone genes(26) and 3. There is a highly non-random codon usage and no discrimination against CG doublets in the coding regions.

DISCUSSION

We have analyzed by DNA sequencing and S1 nuclease analysis four of the histone genes contained in the genomic clone MM221. The two H3 genes which we have identified code for known H3 proteins whereas the H2b and H2a genes code for similar but different proteins than have previously been identified by pro-

tein sequencing. S1 analysis reveals that these genes all code for a small minority (3-5%) of each type of messenger RNA which share enough homology to these genes to be resistant to S-1 digestion. These genes are expressed to the same extent in other mouse cells we have tested(Graves and Marzluff,unpublished results). It is possible too that the protein products of the H2b and H2a genes comprise even a smaller percentage of the population than do their transcripts, since other histone genes which do not share enough homology to be detected by S1 nuclease mapping could exist in mouse myeloma cells. It therefore is not surprising that these H2b and H2a gene products have not been reported in the protein sequence data reported.

It has been estimated by Jacob (33) that there are 10-20 copies of each type of histone gene in the mouse and this number is in line with the estimated copy number of histone genes in other higher eukaryotes (34,35). We have observed that upon Southern analysis, the H3.2 gene hybridizes to 10 other DNA fragments containing about 15 H3 genes (unpublished results). Since the H3.1 and H3.2 genes from MM221 apparently code for a low percentage of the mRNAs yet H3.1 and H3.2 histones are abundant histones we presume that there are more genes which could code for these proteins. We have recently isolated several other histone genes which are expressed in these cells (Graves and Wellman, unpublished results).

There are very few sequence homologies in the 5' and 3' flanking sequences among the four genes from MM221 and the mouse H4 gene isolated by Seiler and Birnstiel (25). The 3' ends have a region of homology among all 5 mouse histone genes sequenced, a 20 base sequence which can form a hairpin loop at the 3' end of the mRNA. This region can form a nearly identical hairpin loop in all five cases (Figure 8b), similar to that found in other histone genes(26). Both our data and the data of Seiler and Birnstiel(25) map the end of the mRNAs near the end of this hairpin. Whether or not this site is the true termination site of transcription is not known. The hairpin structure may only provide a site of secondary structure in the mRNA which is recognized for processing of a larger transcript. The lack of homology found among the 5' ends of the histone genes in this cluster

is not dramatically different from the situation in other histone genes. Only small regions of homology are evident in the 5'ends of the early sea urchin genes(26,32) and in the H2b genes in the chicken(31).

We have also determined the starting point of the mRNAs at the 5' end by S1 nuclease mapping. Each mRNA starts 20-25 nucleotides from an AT rich sequence (TATAA box) similar to that at the 5' end of other eucaryotic genes. Other than this sequence there are only short regions of tenuous homology among the 5' ends of the mouse histone genes. In particular there is no detectable homology in the two H3 genes which are very similar in the coding regions. The short 5' untranslated regions of the mouse histone mRNAs share only the common property of being pyrimidine rich.

The codon usage of the mouse histone genes is clearly not random. There is a preference in all the mouse histone genes for either a G or C in the third base position which helps account for the high GC content(65%) of the coding regions. A consequence of this codon usage is a high frequency of CG doublets, suggesting that selective methylation of these genes is not important in their regulation.

All of the genes from MM221 are controlled in parallel. The synthesis rate and stability of these mRNAs is tightly linked to deoxynucleotide metabolism(20). There must be common features among the mRNA structures to control the rate of mRNA degradation and in the gene structure to regulate the rate of transcription. We are currently trying to define these structures.

ACKNOWLEDGEMENTS

This work was supported by NIH grant GM 29832 to W.F.M. D.B.S. was supported by NIH fellowship GM 07158. We thank Al McGraw, David Brown and Cindy Sprecher for technical assistance.

*Present address: Department of Biochemistry, University of Mississippi Medical Center, 2500 N. State Street, Jackson, MS 39216, USA

REFERENCES

1. Kedes, L. H., (1979) Ann. Rev. Biochem. 448 837-70.
2. Lifton, R. P., Goldberg, M. S. Karp, R.W., Hogness, D. S.

(1977) Cold Spring Harbor Symp. Quant. Biol. 42 1045-51.
3. Hereford, L. M., Fahrner, K., Woolford, J., Rosbash, M. and Kahlback, D. (1979) Cell 18 1261-71.
4. Engel, J. D. and Dodgson, J. B. (1981) Proc. Nat. Acad. Sci. 78 2856-60.
5. Sittman, D. B., Chiu, I. M., Pan, C. J., Cohn, R. H., Kedes, L. H. and Marzluff, W. F. (1981) Proc. Nat. Acad. Sci. 78 4078-82.
6. Heintz, N., Zernik, M. and Roeder, R. G. (1981)Cell 24 61-68.
7. Zernik, M., Heintz, N., Boime, I. and Roeder,R. G. (1980) Cell 22 807-15.
8. Childs, G., Nocente-McGrath, C., Lieber, T., Holt, C. and Knowles, J. A. (1982) Cell 31 383-93.
9. Maxson, R., Mohun, T. Gormezano, G., Childs, G. and Kedes, L. H. (1983) Nature 301 120-26.
10. Isenberg, I. (1979) Ann. Rev. Biochem. 48 159-91.
11. Franklin, S. G. and Zweidler, A. (1977) Nature 266 273-75.
12. Chang, A. C. Y. and Cohen, S. N. (1978) J. Bacteriol. 134 1141-1147.
13. Viera, J. and Messing, J. (1982) Gene 19 269-76.
14. Maxam, A. M. and Gilbert, W. (1977) PNAS 74 560-564.
15. Maxam, A. M. and Gilbert, W. (1980) Methods in Enzymology 65 499.
16. Rubin, C. M. and Schmid, C. W. (1980) Nuc. Acids Res. 8 4613-19.
17. McGraw, R. (1982) Ph.D. thesis, Florida State Univ.
18. Casey, J. and Davidson, N. (1977) Nuc. Acids Res. 4 1539-52.0
19. Weaver, R. and Weissman, C. (1979) Nuc. Acids Res. 7 1175-93.
20. Sittman,D. B., Graves, R. A. and Marzluff, W. F. (1983) Proc. Nat. Acad. Sci. 80 1849-54.
21. Maniatis, T., Hardison, R. C., Lacy, E., Laver, J., O'Connell, C., Quon, D., Sim, G.K. and Efstratiadis, A. (1978) Cell 15 687-701.
22. Manley, J. L., Sharp, P. A. and Gefter, M. L. (1979) J. Mol. Biol. 135 171-97.
23. Green, M.R. and Roeder, R.G.(1980) Nuc. Acids Res. 22 231-42.
24. Hentschel, C., Irminger, J.-C., Bucher, P., and Birnstiel, M.L. (1980) Nature 285 147-151.
25. Seiler-Tuyns, A. and Birnstiel, M. L. (1981) J. Mol. Biol. 151 607-25.
26. Birnstiel, M. L. and Hentschel, C. C. (1981) Cell 25 301-05.
27. D'Andrea, R., Harvey, R. P. and Wells, J. R. E. (1981) Nuc. Acids. Res 9 3119-28.
28. Modiano,G., Battistuzzi,G. and Motulsky, A.G.(1981) Proc. Nat. Acad. Sci. 78 1110-14.
29. Dugaiczyk,A., Law, S.W., and Dennison,O.(1982) Proc. Nat. Acad. Sci.79 71-75.
30. Grandy, G.K., Engel, J.D. and Dodgson, J.B.(1982) J. Biol. Chem. 257 8577-80.
31. Harvey, R.P., Robins, A.J. and Wells, J.R.E.(1982) Nuc. Acids Res. 10 7851-63.
32. Sures, I., Lowry, J. and Kedes, L.H.(1978) Cell 15 1033-44.
33. Jacob, E. (1976) Eur. J. Biochem. 75 275-84.
34. Jacob, E., Malacinsky, G., and Birnstiel, M. L. (1976) Eur. J. Biochem. 69 45-54.
35. Wilson, M. C., Melli, M. and Birnstiel, M. L. (1974) Biochem. Biophys. Res. Comm. 61 404-411.