Volume 8 Number 5 1980 Nucleic Acids Research

# Ubiquitous and gene-specific regulatory 5' sequences in a sea urchin histone DNA clone coding for histone protein variants

M.Busslinger, R.Portmann, J.C.Irminger and M.L.Birnstiel

Institut für Molekularbiologie II, Universität Zürich, Honggerberg, 8093 Zürich, Switzerland

Received 2 January 1980

## ABSTRACT

The DNA sequences of the entire structural H4, H3, H2A and H2B genes and of their 5' flanking regions have been determined in the histone DNA clone h19 of the sea urchin Psammechinus miliaris. In clone h19 the polarity of transcription and the relative arrangement of the histone genes is identical to that in clone h22 of the same species. The histone proteins encoded by h19 DNA differ in their primary structure from those encoded by clone h22 and have been compared to histone protein sequences of other sea urchin species as well as other eukaryotes. A comparative analysis of the 5' flanking DNA sequences of the structural histone genes in both clones revealed four ubiquitous sequence motifs; a pentameric element GATCC, followed at short distance by the Hogness box GTATAAATAG, a conserved sequence PyCATTCPu, in or near which the 5' ends of the mRNAs map in h22 DNA and lastly a sequence A, containing the initiation codon. These sequences are also found, sometimes in modified version, in front of other eukaryotic genes transcribed by polymerase II. When prelude sequences of isocoding histone genes in clone h19 and h22 are compared areas of homology are seen to extend beyond the ubiquitous sequence motifs towards the divergent AT-rich spacer and terminate between approximately 140 and 240 nucleotides away from the structural gene. These prelude regions contain quite large conservative sequence blocks which are specific for each type of histone genes.

## INTRODUCTION

It can well be envisaged that a whole range of regulatory processes exists controlling the expression of the sea urchin histone genes, none of which we are as yet equipped to understand at the molecular level. One of the first steps towards developing such an understanding is the identification of regulatory sequences within the DNA.

H1 proteins occur in 1/2 molar concentrations relative to other histones in chromatin (1,2), yet each histone gene occurs once per basic histone DNA repeat unit. It follows that there must be regulation of H1 protein levels, either during transcription, translation or at a post-transcriptional level.

The appearance of histone mRNAs in the cytoplasm and the production of histones is clearly regulated during the cell cycle of rapidly dividing cells (3). A third type of gene control involves the developmental regulation of histone production (4,5). There is now conclusive evidence in the sea urchin that the histones of an individual are not all the same and that special tissues may have histone variants of distinct primary structure (6).

During evolution, divergent subsets of structural genes coding for histones have accumulated and these exhibit mutations of at least two kinds; those leading to amino-acid substitutions which are selectively neutral or advantageous and those mutations leading to synonymous codons which have no direct effect on amino-acid composition. Using simple restriction analysis to screen a large number of λ recombinants containing sea urchin histone DNA (7), we were able to subdivide the available clones into a very large family consisting of the λh22 type (8-11) and smaller families with disparate and unique restriction patterns. An example of such a small family is clone h19.

The DNA sequences lying directly upstream and downstream of the structural gene are of particular interest since they would be expected to harbor regulatory sequences. The 5' and 3' regions adjacent to all five histone genes for both clones h22 and h19 have now been sequenced and this provides us with the unique possibility of identifying, by sequence comparison, conservative and thus potentially interesting DNA sequences. This work has already provided very useful guidelines for the manipulation of gene units (12).

Previously a comparison of the 3' sequences has led to the discovery that the histone genes of sea urchins share a palindromic DNA sequence and an adjacent purine rich DNA segment (13) which together might function as a terminator of transcription or as a processing signal (13,14). Here we show from DNA sequencing experiments that clone h19 codes for histones whose primary structure is at variance with those encoded by clone h22. Furthermore, we identify conserved 5' sequences of Psammechinus histone genes: 1) By comparing the prelude sequences of the H4, H3, H2A and H2B genes from both h19 and h22 we find ubiquitous sequence motifs, that is, sequences held in common by all these genes that are also found in other eukaryotic genes served by polymerase II. 2) By comparing the prelude sequences of isocoding histone genes in clone h19 and h22 we find that further upstream from the ubiquitous sequences there are conservative, some-

times quite large, sequence blocks which are specific for each type of histone gene.

## MATERIALS AND METHODS

### Material

Restriction endonucleases were purchased from New England Bio Labs, bacterial alkaline phosphatase from Worthington and $T_4$ polynucleotide kinase from P-L Biochemicals. ($\gamma^{32}$P)ATP (specific activity: 1000-3000 Ci/mMol) was obtained from Amersham-Searle. Pre-swollen DEAE-cellulose powder was from Whatman, Sephadex G 75 from Pharmacia, agarose from Sigma, acrylamide and bisacrylamide from Serva, hydrazine from Eastman Kodak and piperidine from Merck.

### Preparation of h19 DNA

λh19 is one of the recombinant λ phages originally obtained by S. Clarkson et al. (15) by digestion of sperm DNA of the sea urchin P.miliaris with Hind III and insertion of the 6 kb fraction into the Hind III site of λ598 (16). In a subsequent step we transferred the h19 insert into the high yield vector λ Sam7 (15). The h19 insert can be isolated from the phage DNA as described for h22 DNA (15,17).

### Preparation of end labelled DNA fragments

The conditions for all endonuclease digestions were those suggested by the enzyme suppliers. DNA fragments were dephosphorylated and terminally labelled according to the procedure of Maxam and Gilbert (18), modified as described by Boseley et al. (19). End labelled DNA fragments were separated on 1% or 2% agarose slab gels by electrophoresis in Loening E buffer (20). The DNA was eluted from the gel either electrophoretically (21) or by diffusion (18). This DNA was further purified on a DEAE cellulose column. Each fragment was then cleaved asymmetrically by digestion with a restriction enzyme and the individually labelled fragments were recovered by gel electrophoresis.

### Partial restriction mapping

Intact h19 DNA was labelled at the Hind III ends and then cleaved by either Bgl I or Sal I, both of which cut h19 DNA only once. The resulting fragments, labelled at one end, were isolated as described above. The recognition sites of 23 restriction enzymes were mapped using the partial restriction mapping procedure of Smith and Birnstiel (22). Partial digestions were optimized by diluting the enzyme and by adding sonicated phage $T_4$ DNA.
10μl samples were taken at various time points from the initial 50μl volume

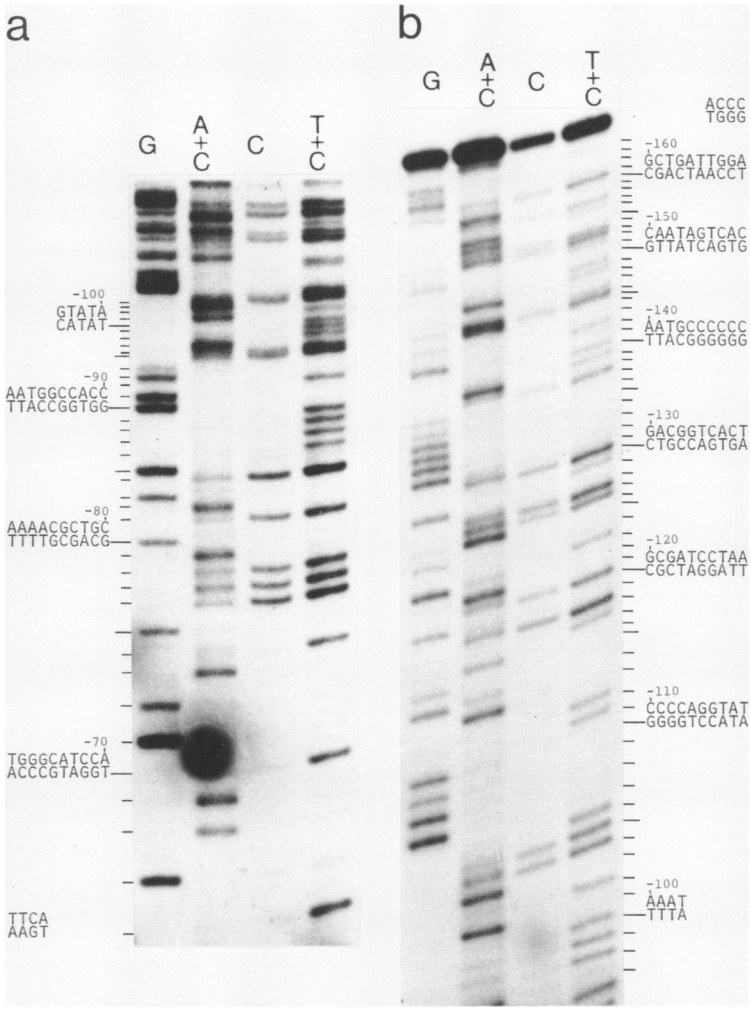


*Fig. 1 a-c Gels showing the prelude sequence of the H2A gene in clone h19. a) and b) Short and long runs of a DNA fragment that was labelled at a Taq I site (-57/-60; see Fig.3) and which contains the sequence motifs PyCATTCPu (-66/-72), GTATAAATGG (-95/-104), GATCC (-114/-118) and the 30 bp long sequence block (-134/-163) found also in clone h22. c) Sequence of the opposite DNA strand labelled at a Hpa II site (-171/-174). The positions of the bases upstream of the initiation codon are indicated by negative numbers (see Fig.3).*

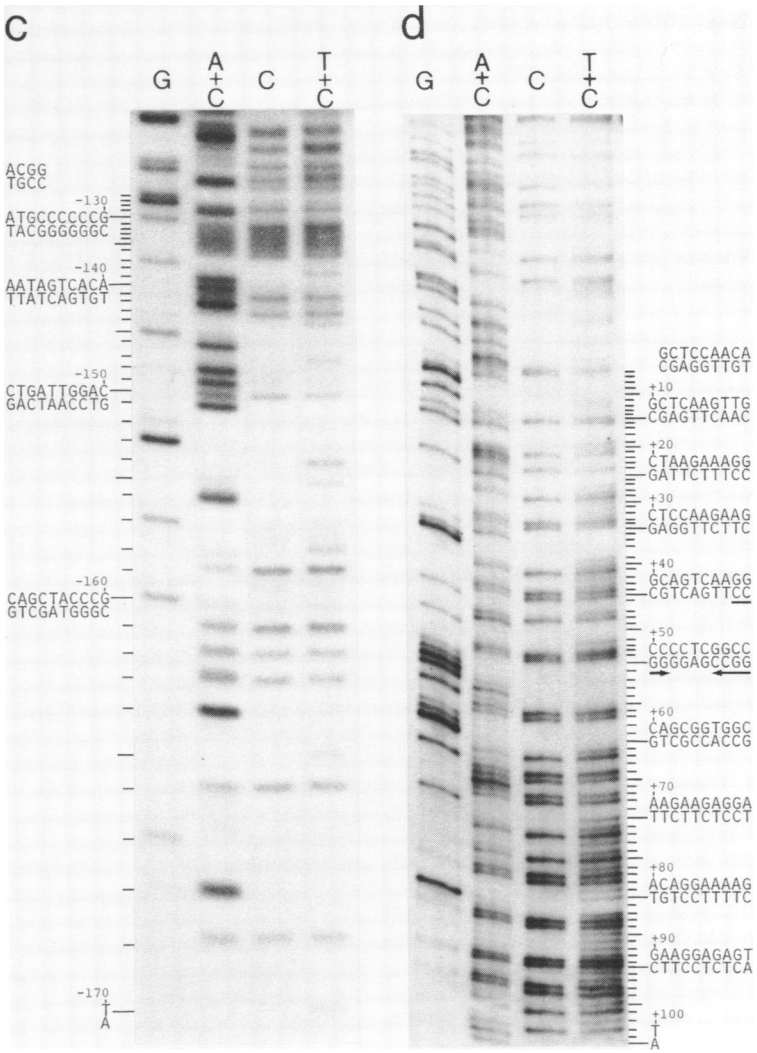


*Fig. 1 d* *Gel showing the DNA sequence coding for the N-terminal part of the H2B gene in clone h19.*
*The sequence was determined from the unique Taq I site near amino acid 67 of the H2B gene. The positions of the bases downstream of the initiation codon are shown by positive numbers. A palindromic sequence giving rise to irregular spacing of bands (+46/+47) in the T-track is indicated by arrows.*

and the reaction was stopped by adding 5µl of a solution containing 0.2% agarose, 20mM EDTA, 10% glycerol. A part of each sample was analyzed by electrophoresis on 1% or 2% agarose gels (40cm x 17cm x 1mm) to determine which time point showed the maximum number of partial digestion fragments. Samples of optimal time points for all analyzed restriction enzymes were run in parallel on so-called master gels allowing the relative positions of all restriction sites to be read directly from the same autoradiogram (22). This information was very helpful in deciding what combination of restriction enzymes would provide the best sequencing strategy.

### DNA sequencing

The DNA sequencing procedure was based on that of Maxam and Gilbert (18) with some modifications already described elsewhere (13). In some experiments restriction sites were used to obtain sequences in both directions, for which overlapping sequence data is not available. In all these instances we verified by partial digestion of a relevant restriction fragment that the site used was indeed unique and that no short DNA segment was missed in the sequencing analysis.

### Computer analysis

DNA sequences were analyzed by computer using a program that presents homologies between two related sequences on diagonals of a matrix, whose axes consist of these two sequences. The diagonal that contained most homologous nucleotides was then chosen to align the two sequences. Homology blocks that were outside of this diagonal could only be accounted for by introducing an insertion or a deletion in one of the two sequences. The optimal alignment was chosen as that which gave the maximum number of homologous base pairs with the minimum number of deletions. In regions of little overall homology the alignment was necessarily arbitrary.

## RESULTS AND DISCUSSION

### Restriction map and gene arrangement for clone h19

h19 DNA is 6.7 kb long. It may be reclaimed from the recombinant λ Sam7 h19 by Hind III digestion and preparative actinomycin-CsCl centrifugation. Using the partial restriction mapping technique of Smith and Birnstiel (22) a detailed map of clone h19 was constructed which is shown in Fig.2. There is very little overall resemblance between the h19 restriction map and that presented earlier for h22 (11). To locate the histone coding sequences in the

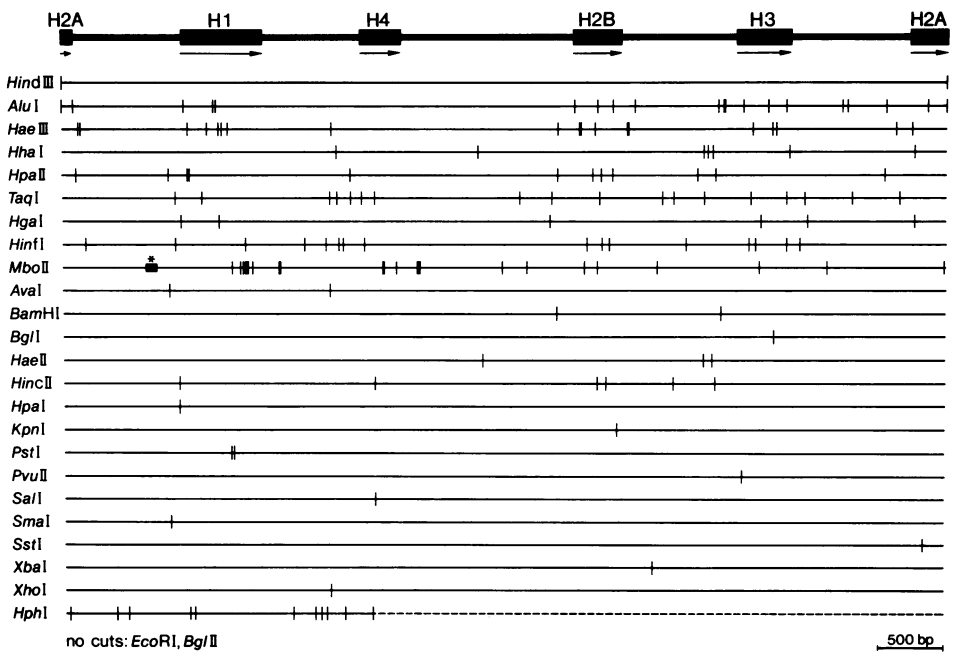


*Fig. 2. Gene arrangement and restriction map of clone h19. The exact locations of the protein coding sequences as well as the polarity of transcription have been determined by DNA sequencing. The restriction mapping was done as described in the Methods section. The sites of Hph I cleavage were not determined in the DNA segment shown by a dashed line. A multiple MboII cut in front of the H1 gene is indicated by an asterisk.*

h19 unit nick-translated DNA fragments, each containing a single histone gene derived from clone h22, were hybridized to different segments of h19 DNA previously immobilized on nitrocellulose filters (23). All coding sequences could be localized, with the exception of the H1 gene which did not cross-hybridize between the two clones.

The exact map positions of all five coding regions including the H1 gene have now been established by DNA sequencing and are shown in Fig.2. The relative arrangement of the histone genes in h19 and h22 is identical, with the difference that in clone h19 the Hind III restriction site falls within the H2A gene while in the h22 repeat it is found in the spacer between the H1 and the H4 gene (11). As in h22 (9), all histone genes share the same polarity and are separated from each other by spacer DNA (10,11).

In Fig. 1 we present typical examples of sequencing gels of the 5'


H4
-380 -370 -360 -350 -340 -330 -320 -310 -300 -290 -280 -270
GAGGGGGGCGGGCGGGCTTGGGTGAGGAATGCATTCCTATTAATGTATGTATATTGAGTAATGTAGGTGATGACCGATTTTTCAGTTAATTAAGTTTCGTCATCCATGTGTACTGTTGGATAAG
-260 -250 -240 -230 -220 -210 -200 -190 -180 -170 -160 -150 -140
ATTCGTCACCCCCTGCTTTTTTCAACTCGAGCATTCATCACAGGCCTCCCTAGATATGCAAGCGAATACTCGCCACAAGGGGGCGCACTCGAATGGGGAGTCTCCGCACTCCAGTCCCGCATACCGAATGAT
-130 -120 -110 -100 -90 -80 -70 -60 -50 -40 -30 -20 -10
GCCGAATCTCGTCACCCAAGTCCGCAATGGTGTAACAATACCCGCTGCAATCCGGTCGAGGCATCATTCGCTTAGCGTAATACCCAGTCTACAGGATCACACAGAACTCGCTCTCAACTATCAATCAGCATC
Ser Gly Arg Gly Lys Gly Gly Lys Gly Leu Gly Lys Gly Gly Ala Lys Arg His Arg Lys Val Leu Arg Asp Asn Ile Gln Gly Ile Thr Lys Pro Ala Ile Arg Arg Leu Ala Arg Arg Gly Gly Val
10 20 30 40
ATGTCAGGTCGAGGAAAAGGAGGAAAGGGACTCGGAAAGGGCGGTGCCAAACGTCATCGCAAGGTTCTACGAGACAACATCCAAGGCATCACCAAGCCTGCAATCCGTCGACTTGCTAGAAGGGGAGGTGTC
Lys Arg Ile Ser Gly Leu Ile Tyr Glu Glu Thr Arg Gly Val Leu Lys Val Phe Leu Glu Asn Val Ile Arg Asp Ala Val Thr Tyr Cys Glu His Ala Lys Arg Lys Thr Val Thr Ala Met Asp Val Val
50 60 70 80
AAGAGGATCTCTGGTCTCATCTACGAAGAGACACGCGGTGTACTGAAGGTCTTCTTGGAGAATGTCATCCGTGATGCAGTCACCTACTGCGAGCACGCCAAGCGAAAGACTGTCACAGCCATGGACGTGGTG
Tyr Ala Leu Lys Arg Gln Gly Arg Thr Leu Tyr Gly Phe Gly Gly
90 100
TATGCACTGAAGAGGCAGGGTCGTACATTGTACGGCTTCGGCGGCTAA
H2B
-380 -370 -360 -350 -340 -330 -320 -310 -300 -290 -280 -270
TCGACATGTTCCATATTCGGATATACTAATTCCCATCAAAATCATCGTCTTCTTCAACTTTTTCATATTTCTTATTCTTTGGGGATTGAATTGAAAATTGCAGAGAGATCCTATATTAGAACAT
-260 -250 -240 -230 -220 -210 -200 -190 -180 -170 -160 -150 -140
AATTATGCCCGTATGATCTGTATGGGTTGTCACTATGTCGCCATCTTTAGGCAGGGGATGGTACAGGCACTAATTCTGCGACGCCTATAGCCCAATGAAAGGATCGAGACCGAGGCTCATTTGCATACGGAC
-130 -120 -110 -100 -90 -80 -70 -60 -50 -40 -30 -20 -10
CGCAGCATACGGATCCGGCCCCGCGTGTATAAAAAGGAATGGTTCTCGCTTGCCATTCACAGTATCCAAAGATTATTTGCTTGACATACTCGTTTCGCTGCATCTTTACAGACCAGAAAACCTCAATTCATC
Ala Pro Thr Ala Gln Val Ala Lys Lys Gly Ser Lys Lys Ala Val Lys Ala Pro Arg Pro Ser Gly Gly Lys Lys Arg Asn Arg Lys Arg Lys Glu Ser Tyr Gly Ile Tyr Ile Tyr Lys Val Leu Lys
10 20 30 40
ATGGCTCCAACAGCTCAAGTTGCTAAGAAAGGCTCCAAGAAGGCAGTCAAGGCCCCTCGGCCCAGCGGTGGCAAGAAGAGGAACAGGAAAAGGAAGGAGAGTTATGGAATCTACATCTACAAAGTCCTCAAG
Gln Val His Pro Asp Thr Gly Ile Ser Ser Arg Ala Met Ile Ile Met Asn Ser Phe Val Asn Asp Ile Phe Glu Arg Ile Ala Gly Glu Ser Ser Arg Leu Ala Gln Tyr Asn Lys Lys Ser Thr Ile Ser
50 60 70 80
CAGGTTCATCCAGATACCGGCATCTCCAGTCGGGCCATGATCATCATGAACAGCTTCGTCAACGACATCTTCGAGCGAATTGCCGGCGAATCTTCCCGCCTCGCTCAGTACAACAAAAAGTCAACCATCAGC
Ser Arg Glu Ile Gln Thr Ala Val Arg Leu Ile Leu Pro Gly Glu Leu Ala Lys His Ala Val Ser Glu Gly Thr Lys Ala Val Thr Lys Tyr Thr Thr Ser Lys
90 100 110 120
AGTCGCGAGATTCAGACCGCCGTCCGCCTCATTCTCCCCGGAGAGCTGGCAAAGCACGCTGTGAGCGAGGGTACCAAGGCAGTGACGAAATACACTACCTCCAAGTAG
H3
-280 -270
AAATTCCCGGCTTGCCATGTAATG
-260 -250 -240 -230 -220 -210 -200 -190 -180 -170 -160 -150 -140
TAACCCATGTTTAAAACGGTGGTAGCGCCACCTTTCGACACATTCTAATGACTTTGCGCATACGGCATCGCCAAGCCCCCCTTCCCGTCAGGCGCTAAACAAAAGAACGACCCGGTTGACCAATCAAGAGAG
-130 -120 -110 -100 -90 -80 -70 -60 -50 -40 -30 -20 -10
CTTTACAAAGGTAGGGCGACGGCACCAGGATCCTTCGCAGCTCATATAAATAGCTGAACATTGCCAGTGGTTCTCATTCATCCCGTCACTCGTATTTGAAGTACCGAATTACCTACTGTCTCTCAAGCAACT
Ala Arg Thr Lys Gln Thr Ala Arg Lys Ser Thr Gly Gly Lys Ala Pro Arg Lys Gln Leu Ala Thr Lys Ala Ala Arg Lys Ser Ala Pro Ala Thr Gly Gly Val Lys Lys Pro His Arg Tyr Arg Pro
10 20 30 40
ATGGCACGCACCAAGCAGACCGCTCGCAAATCTACAGGAGGGAAGGCTCCCCGCAAGCAGCTGGCAACCAAAGCAGCCAGAAAGAGTGCCCCCGCCACTGGAGGAGTCAAGAAGCCTCATCGATACAGGCCT
Gly Thr Val Ala Leu Arg Glu Ile Arg Arg Tyr Gln Lys Ser Thr Glu Leu Leu Ile Arg Lys Leu Pro Phe Gln Arg Leu Val Arg Glu Ile Ala Gln Asp Phe Lys Thr Glu Leu Arg Phe Gln Ser Ser
50 60 70 80
GGCACAGTCGCCCTGAGAGAGATTCGCCGCTACCAGAAGAGCACCGAACTTCTCATCCGAAAACTGCCATTCCAGCGTCTAGTGCGTGAGATTGCACAGGACTTCAAGACAGAGCTACGTTTCCAGAGTTCC
Ala Val Met Ala Leu Gln Glu Ala Ser Glu Ala Tyr Leu Val Gly Leu Phe Glu Asp Thr Asn Leu Cys Ala Ile His Ala Lys Arg Val Thr Ile Met Pro Lys Asp Ile Gln Leu Ala Arg Arg Ile Arg
90 100 110 120 130
GCTGTGATGGCCCTTCAAGAAGCCAGCGAGGCATACCTGGTTGGCCTCTTTGAGGATACCAACCTGTGTGCCATCCACGCCAAGAGGGTTACCATCATGCCCAAAGACATCCAGCTCGCCCGTCGAATCCGC
Gly Glu Arg Ala
GGAGAACGCGCCTAG

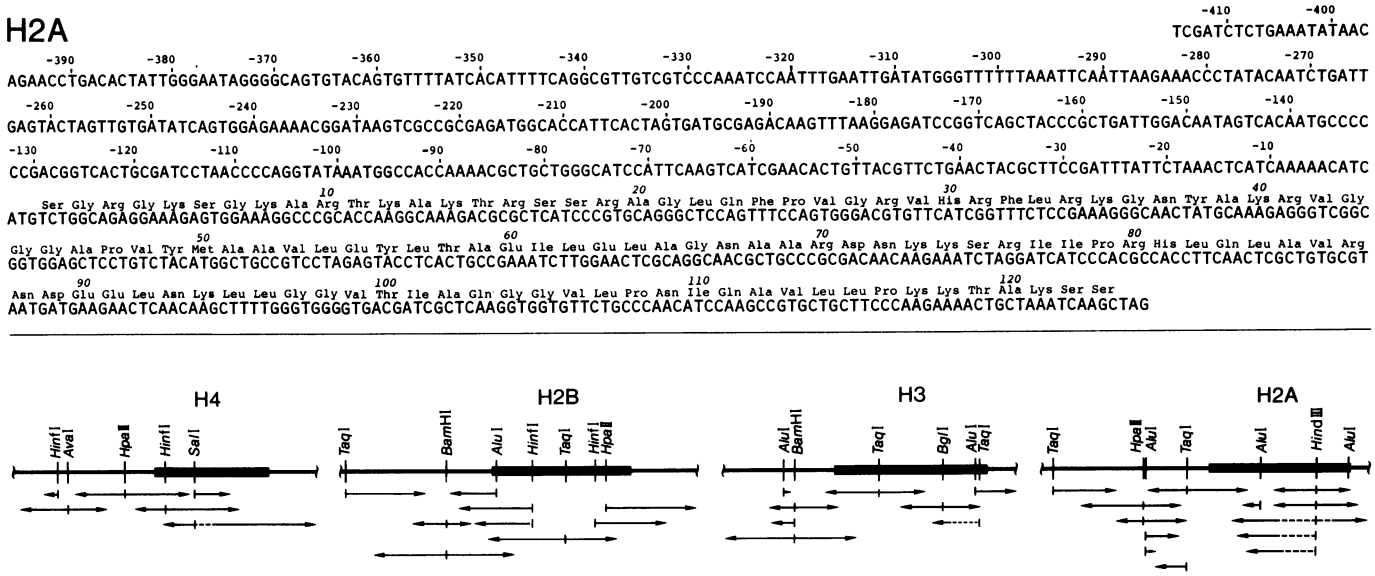


*Fig. 3. DNA sequences of the structural H4, H2B, H3 and H2A genes and their 5' flanking regions in clone hl9. Only the DNA strand having the same polarity as the mRNA is shown. The sequencing strategy is depicted below (see also Methods section). Each arrow represents the length of nucleotide sequence analyzed in an individual experiment and starts at the restriction site labelled for sequencing. Note that the amino-acids of the structural genes are numbered in italics and the nucleotides in the 5' flanking sequences are indicated by negative integers, relative to the initiation codons.*

region of the H2A gene. Included is a sequencing gel of the N-terminal part of the structural H2B gene, showing a very GC-rich palindromic DNA sequence that gives rise to irregular spacing of some bands in the T-track. Since this GC-rich region is prone to give rise to sequencing artefacts the complementary DNA strand was sequenced in the opposite direction, as previously described for an analogous situation (13). The 5' flanking, as well as the protein coding DNA sequences of the H4, H2B, H3 and H2A genes are presented in Fig.3 together with the derivation of the amino-acid sequences where appropriate. Only the DNA strand having the same polarity as the mRNA is shown. The strategy and the restriction sites used for sequencing are depicted at the bottom of Fig.3. For the presentation of DNA sequences the 5' terminal part of the H2A gene to the left of the HindIII site (see Fig.2) is joined to the 3' terminal part to the right of the HindIII site. This joining generates a typical H2A protein sequence across the HindIII site without any disturbance in the reading frame, which suggests that at least three h19 units are linked together on the chromosomal DNA of Psammechinus miliaris. Quantitative hybridization studies indicate that h19 is repeated only a few times (ca.5 times) in the haploid genome of the sea urchin Psammechinus. Thus, h19 constitutes a separate mini-family of histone genes. The h22 unit, on the other hand, has a reiteration frequency two orders of magnitude greater than h19 and, as a maxi-family, is a representative of the most common histone DNA unit of Psammechinus (7).

h19 DNA segments coding for H4, H3, H2A and H2B proteins

The entire structural H4, H3, H2A and H2B genes have been sequenced. The coding regions were identified by reading the DNA sequences in three frames and by comparing the resulting amino-acid sequences with available protein data. Such comparative analysis revealed that the H4 amino-acid sequence encoded in clone h19 is identical to the H4 protein isolated from gonads of the same sea urchin P.miliaris (24), but that it differs from calf thymus H4 (25,26) at position 73 (Cys↔Thr). The H3 amino-acid sequence encoded by clone h19 has a Glu at position 81 while H3 proteins isolated from calf thymus (27), carp testis (28), shark erythrocytes (29) and chicken erythrocytes (30) contain an Asp at this position. There is a further exchange at position 96 (Ser ⟷ Cys) between the amino-acid sequence of h19 and calf thymus H3.

A comparison between H2A and H2B protein sequences and their counter-

parts coded by clone h19 is shown in Fig.4. The H2A encoded by clone h19 differs from a gonadal H2A of the same sea urchin P.miliaris (31) in 15 positions, involving conservative as well as non-conservative amino-acid exchanges. A similar degree of amino-acid substitution is seen between the H2A encoded by clone h19 and the H2A's of terminally differentiated tissues of vertebrates (32,33) or some transformed vertebrate cells (34,35). The central part of the H2A of h19 (amino-acid residue 50-74) is identical to a partial amino-acid sequence of an H2A variant isolated from $18^h$ embryos (6) of Parechinus angulosus, a sea urchin of the same family as P.miliaris (36). This suggests that h19 might code for embryonic histone variants. This becomes more evident when the N-terminal part of the H2B of clone h19 is compared with that of a partially sequenced H2B variant of the same embryonic stage ($18^h$) of P.angulosus (37). Both proteins are identical in the first 27 amino-acids except for a conservative substitution (Ala⟷Gly) at position 4. In marked contrast to this, the H2B of clone h19 shows little homology in the N-terminal part to three H2B variants isolated from sperm of P.angulosus (38), nor does it show much homology with H2B proteins isolated from adult tissues of other animals (39-43). It is only in the central and the C-terminal part that these proteins exhibit a degree of homology comparable to that seen for the H2A proteins.

Table I summarizes the result of a comparative analysis of the amino-acid sequences derived from the DNA sequences of four sea urchin histone DNA clones (h19, h22, pSp2, pSp17). Together, clone pSp2 and pSp17 code for all five histones (44-46) and are representatives of the majority type of histone DNA repeats in Strongylocentrotus purpuratus (47,48). No amino-acid substitution can be found for the H3 and H4 proteins in all four clones. The H2A proteins of clone h19 and clone h22 differ by two amino-acids. The H2B proteins of these two clones show ten amino-acid substitutions, six of which are non-conservative (position 17,20,57,74,78,121). The H1 protein sequences deduced from DNA sequences of both clones suggest that these two proteins differ by 30% (manuscript in preparation). Thus clone h19 and clone h22 code for distinct H2A, H2B and H1 proteins and have therefore been separated from one another in evolution for some time. The relatively small incidence of amino-acid substitutions fixed in the course of evolution is in marked contrast to the much larger incidence of mutations to synonymous codons, giving rise to an average DNA sequence divergence of 12,4% in all

## H2A proteins

```
                                10        20        30        40        50        60        70        80        90       100       110       120
clone h19                SGRGKS-GKARTKAKTRSSRAGLQFPVGRVHRFLRKGNYAKRVGGGAPVYMAAVLEYLTAEILELAGNAARDNKKSRIIPRHLQLAVRNDEELNKLLGGVTIAQGGVLPNIQAVLLPKKTA-KSS
gonad of P. miliaris (31)         G-A GKA  S                  N  A    L      A          T        I                                  GS  K
calf thymus (32)                  QG   A                    L  E  A    L                 T        I             K                    ESHHKAKGK
mouse friend leukemia cell (34)   EG   A  S                 L  E  A                      T        I             K   E       E        ESHHKAKGK
rat chloroleukemia cell (35)      QG   A  S                 L  E  A    L                 T        I             K                    ESHHKAKGK
chicken erythrocyte (33)          QG   A  S                 L  E  A    L                 T        I             K                    DSHKAKAK
```

## H2B proteins

```
                                10        20        30        40        50        60        70        80        90       100       110       120
clone h19                APTAQVAKKGSKKAVKAPRPSGGKKRNRKRKESYGIYIYKVLKQVHPDTGISSRAMIIMNSFVNDIFERIAGESSRLAQYNKKSTISSREIQTAVRLILPGELAKHAVSEGTKAVTKYTTSK
sea urchin sperm
of P. angulosus (38)
1      PSQKSPTKRSPTKRSPTKRSPQKGGKGGKGAKRGGKAGKRRRGVQV  R R R                    SV       V     A AG  TT  RR  V    V      L                  R
2       PRSPAKTSPRKGSPRKGSPSRKASPKRGGKGAKRAGKGGRRRRVV  R R R                    SV       V       A   TSA RR  V           L                  R
3  PRSPAKTSPRKGSPRKGSPSRKASPKRGGKGAKRAGKGGRRRRVV  R R R                    SV       V     S A   TSA RR  V           L                  R
trout testis (39)                      PEPAKSAPKKGSKKAVTKTAGKGGK  K S     A  V              K  G                    H   R   T         L                  S
patella testis (40)                     PPKVSSKGAKKAGKAKAARSGDK  K R     S              V  K  S                 A A  H   R   T         L                  S
drosophila (cf 41)                      PPKTSGKAAKKAGKAQKNITKTDK KK      A                 K  S                 A A  H   R   T         L                  S
calf thymus (42) and                   PEPAKSAPAPKKGSKKAVTKAQKKDGK  K S     SV V              K  G                   A   H   R   T         L                  S
human spleen (43)
```

*Fig. 4. Comparison of known H2A and H2B protein sequences with their counterparts of clone h19. Only amino-acid substitutions relative to the sequence of clone h19 are indicated in the case of the H2A proteins and in the constant part of the H2B proteins (residue 25-122). Amino-acids in the variable region of the H2B proteins (up to position 24) are shown in full. No attempts have been made to align them for homology. The above numbering refers to the amino-acid sequence of clone h19. The one letter code for amino-acids is: 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.*

Table I Comparison of amino acid sequences coded by four sea urchin histone DNA clones

| histone | amino acid position in clone h19 | amino acid substitution | compared histone DNA clones |
|---|---|---|---|
| H4 | - | None | h19 ↔ h22 (11) |
| | - | None | h19 ↔ pSp2 (46) |
| H3 | - | None | h19 ↔ h22 (11) |
| | - | None | h19 ↔ pSp17 (45) |
| H2A | 15 | Thr ↔ Ser | h19 ↔ h22 (11) |
| | 120 | Ala ↔ Gly | |
| | - | None | h19 ↔ pSp17 (45) |
| H2B | 4 | Ala ↔ Gly | h19 ↔ h22 (11 and unpublished results) |
| | 17 | Ala ↔ Pro | |
| | 20 | Pro ↔ Ala | |
| | 27 | Asn ↔ His | |
| | 51 | Ile ↔ Val | |
| | 57 | Ile ↔ Thr | |
| | 74 | Ser ↔ Ala | |
| | 78 | Ala ↔ Thr | |
| | 98 | Ile ↔ Leu | |
| | 121 | Ser ↔ Ala | |
| | 17-21 | Ala Pro Arg Pro Ser ↔ Gly Thr Lys Thr Ala X | h19 ↔ pSp2 (45) |
| | 57 | Ile ↔ Val | |

four structural genes.

Most strikingly, the clone h19 belonging to a mini-family of histone variants in Psammechinus miliaris, and the clones pSp2 and pSp17 belonging to the majority class of histone genes in S.purpuratus, code for identical H2A and nearly identical H2B protein sequences, with only one conservative amino-acid exchange at position 57 and an apparent mutational hot spot in the N-terminal part of the H2B protein (amino-acid residue 17-21; for corresponding DNA sequence see Fig.1d). Even more astonishingly, these histone DNA clones (h19 and pSp2/pSp17) show an average divergence of only 1,74% in the DNA sequences coding for H4, H3, H2A and H2B proteins. Thus it is apparent that the DNA sequence divergence between h19 and h22 (12,4%), two clones from the species P.miliaris, is far greater than that between h19 and the S.purpuratus clones pSp2 and pSp17 (1,74%). This indicates that the evolutionary relationship between h19 and pSp2/pSp17 appears to be

orthologous (50), i.e. their DNA sequence divergence reflects the time passed since the speciation of the two sea urchins. h22 and pSp2/pSp17, representing maxi-families coding for the bulk of cleavage stage histone mRNAs in both species (14,48,49), are then paralogous DNA sequences (50), i.e. they diverged long before the two species separated. This concept and its implications for the mode of histone gene evolution and possible significance of histone variants in development will be discussed in detail elsewhere (manuscript in preparation). Since clone h19 and h22 have clearly diverged in evolution, we can now use the sequencing data to identify conservative and hence potentially interesting DNA sequences.

Ubiquitous conserved 5' prelude sequences

Previous work has identified a TATAAATA motif, the so-called Hogness box, in the prelude sequences of many eukaryotic genes (51) and the similarity of this sequence to the Pribnow box has been noted (51). Homologies between the 5' flanking regions of the structural histone genes in h19 and h22 have been compiled in Fig.5. The comparison of these prelude sequences reveals at least four areas of sequence homology, one of which is a heptameric consensus sequence PyCATTCPu. Sea urchin histone mRNAs are capped, most likely involving a terminal A site (60,61). For h22, S1 mapping has placed the 5' ends of histone mRNAs in or near the sequence PyCATTCPu and hence this motif has been referred to as the "cap sequence" (14).

Some 16-24 nucleotides further upstream in the H2A, H2B and H3 preludes there is the decameric consensus sequence GTATAAATAG, which 8-10 nucleotides upstream is preceded by the conserved pentameric element GATCC. Both these motifs are found, in slightly modified form, at corresponding positions in the H4 preludes. Other eukaryotic genes have similar conserved sequences to those found in histone genes (Fig.5). Quite possibly therefore our compilation reveals ubiquitous sequences shared by other genes which are transcribed by polymerase II.

The DNA regions encoding the histone mRNA leaders, downstream from the "cap sequence", show a non-random nucleotide pattern manifested by a strong under-representation of the nucleotide G and terminate with the sequence $\overset{C}{A}A\bar{P}y\dot{C}\dot{A}TG$ in the histone genes, as in many eukaryotic mRNA genes sequenced to date (see Fig.5). This sequence containing the first initiation codon in the leader of histone mRNAs might be a signal for the initiation of translation. In addition there is a pyrimidine-rich homologous sequence, 12 base


*Fig. 5. Ubiquitous sequences upstream of the structural histone genes in clone h19 and h22. h19 DNA sequences were compared to those of h22 (11; and unpublished results). The sequence elements GATCC and GTATAAATAG, common to the H3, H2A and H2B preludes, are found in modified form at corresponding positions in the H4 preludes. DNA sequences of other eukaryotic genes have been aligned in the lower part of the figure. The position of the 5' end of the mRNA in these genes is marked by an asterisk. The alignment of the yeast iso-1-cytochrome c sequence is tentative. N.D. denotes undetermined bases.*

pairs long, some 6 to 16 nucleotides upstream from the ATG of the H3 and H4 genes (7).

Gene-specific conserved 5' sequences in clone h19 and clone h22

In Fig. 6 we attempt to show the evolutionary pattern of larger areas of sea urchin histone DNA by matching the preludes of isocoding histone genes analyzed in clone h19 and h22. To highlight both these homologies and deletion/insertion events we have boxed in the former and have indicated the latter by black bars. Taking base substitutions and deletions into account, areas of homology are seen to extend upstream of the 5' end of the mRNA and of the ubiquitous Hogness box towards the divergent AT-rich spacer, terminating between approximately 140 (H4 prelude) and 240 (H2B prelude) nucleotides away from the structural gene. This pattern of sequence conservation is seen most strikingly in the case of the H2A gene region. Note that the homologies upstream of the consensus sequence GATCC are confined to isocoding genes and hence are gene-specific.

The most impressive homology block within the prelude regions is found 140 base pairs upstream of the structural H2A gene. It consists of a continuous stretch of 30 nucleotides (at position -144 to -173 (Fig.6)) interrupted by only one mismatched base at its center. The central portion of the homology block is AT-rich, in contrast to the GC-rich flanking regions. It contains a repeated pentanucleotide ACAAT and a region of dyad symmetry ATTG GACAA(T/A)TTGTCACAAT. In bacterial systems such dyad symmetries are often diagnostic of protein-DNA interactions, as for example with restriction enzymes, repressor proteins and the catabolite gene activator protein (62).

Previously, the existence of prelude regions in histone DNA was inferred from partial denaturation mapping of h22 DNA (10). Later, they were recognized as regions of overall high GC content by DNA sequence analysis and were therefore found to be quite distinct from the AT-rich spacer DNA (11). Here, we show that prelude regions are also unique entities and can be distinguished from spacer, in that important sections within them are conserved in evolution. This suggests that they may play a role in the processes of histone gene expression. Thus the histone DNA of the sea urchin presents itself as a series of five structural genes with conserved 5' and 3' flanking sequences separated from one another by less conserved spacer DNA.

Possible biological significance of the conserved sequence elements

The promoter of prokaryotic genes, on the basis of sequence homologies, can

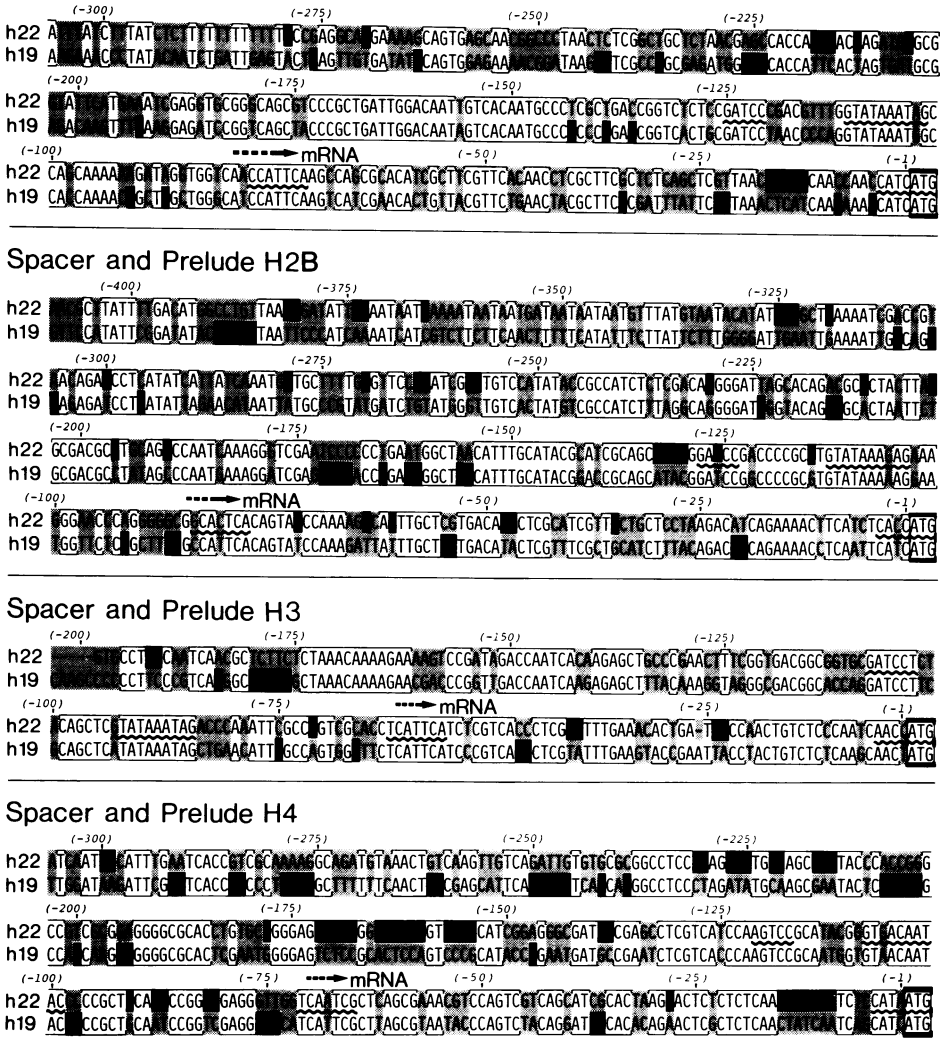


*Fig. 6. Gene-specific conserved 5' sequences in clone h19 and h22. Prelude sequences of isocoding histone genes in clone h19 and h22 (11; and unpublished results) have been aligned for maximal homology, as described in the Methods section. Insertion/deletion events are indicated by black bars, divergent sequences are shaded and homology blocks are bracketed. The ubiquitous sequence motifs (see Fig.5) are shown by wavy lines. For h22, the approximate map positions of the 5' ends of histone mRNAs are indicated by arrows. Note that the numbering in this figure does not refer to the absolute nucleotide positions, but rather is a measure of distance in this new alignment.*

be ordered into the Pribnow box, proximal to the 5' end of the mRNA coding sequences (63,64), and the entry site for the RNA polymerase distal to it (62). Using phosphoethylated promoter DNA, W. Gilbert has demonstrated that the contact sites between repressor protein and DNA, and between catabolite gene activator protein and DNA, all map on one side of consecutive turns of the double helix (65, and personal communication). This is only possible if the distances between such recognition signals are rigidly maintained. By contrast the distance between the Pribnow box and the site of initiation of transcription shows considerably less topological constraints (H. Schaller, personal communication), possibly because the DNA sequence is denatured during promotion (66).

We now have extensive comparable data on the 5' sequences of at least ten histone genes, eight of them presented, so that it is possible to examine the arrangement of conserved sequences within these regions in the light of the prokaryotic pattern of sequence conservation. The length variation of DNA sequences separating the conserved, ubiquitous DNA signals of the eight histone genes is portrayed in Fig.5, where the "missing" bases have been accentuated by a black overlay. The GATCC and the GTATAAATAG motifs are separated from one another by a more or less constant eight to ten nucleotides. It could be easily imagined that a regulatory factor might act by making contact with these two conserved sequences in a defined stereochemical way. It has been suggested that the distance between the Hogness box and the sequence coding for the 5' terminus of several eukaryotic mRNAs is constant (51,53,54; see Fig.5) and that these two sequences are compulsorily linked with one another, as far as their helical topologies are concerned (53,54). On the other hand, the interval between Hogness box and "cap sequence" in the histone genes can vary by as much as eight nucleotides, i.e. nearly one turn of the DNA helix. If the PyCATTCPu motif were in fact the sequence coding for the 5' end of histone mRNAs then there would obviously be a basic conflict with the situation found in other eukaryotic mRNA genes. This question can only be resolved once the 5' termini of histone mRNAs are sequenced and can therefore be accurately mapped.

Furthermore, it is of interest to analyze the distribution of small "sequence deletions" between conserved prelude sequences of isocoding genes. As can be seen in Fig.6, such "deletions" map mainly between the initiation codon and the Hogness box whereas the distance between the Hogness box and

the sequence GATCC is rather constant. In the region immediately upstream of the GATCC pentamer no deletions are found in the H3 prelude and only three single-base deletions in the case of the H2A prelude. Since extensive base deletions are not permitted in these areas it would appear that helical topologies between gene-specific homology blocks of the H3 and H2A preludes are conserved. On the other hand the corresponding regions of the H2B and H4 genes have apparently not been under such high evolutionary constraints.

Our comparative sequencing studies have revealed the existence of conserved sequence motifs, some of which are shared with other genes transcribed by polymerase II while other motifs are specific for each type of histone gene. There is already a considerable body of evidence (12) that deletions of the conserved sequences in the prelude of the H2A gene elicit both qualitative and quantitative changes in the expression of this gene. For instance when the H2A gene unit, modified by deletion of the H2A-specific conserved sequence block (in Fig.6, position -144 to -173), is introduced into the frog oocyte nucleus, transcription of the H2A gene is accelerated. When the area containing the Hogness box and the GATCC motif is deleted, the rate of transcription if reduced and a family of H2A mRNAs containing erroneous 5' termini is generated. Deletion of the ubiquitous "cap sequence" has led to the production of a truncated H2A mRNA which could still be translated in the frog oocyte (E. Probst, unpublished).Thus analyzing a cloned gene unit in terms of the evolutionary conservation of sequences has enabled us to locate sequences which, when tested in surrogate genetics experiments, have proved to have specific regulatory functions.

ACKNOWLEDGEMENTS

We wish to thank Heidi Schernitzki for excellent technical assistance, Fritz Ochsenbein for graphical work and Silvia Oberholzer for typing the manuscript. We are grateful to Dr. M. Chipchase for critical reading of the manuscript. This work was supported by a grant of the State of Zürich and by the Swiss National Research Council, grant No.3.066.076 and No.3.257.077.

REFERENCES

1. Kornberg, R.D. (1974) Science 184, 868-871.
2. Goodwin, G.H., Nicolas, R.H. and Johns, E.W. (1977) Biochem.J. 167,485-488.
3. Borun, T.W. (1975) Histones, Differentiation and the Cell Cycle, in *Cell Cycle and Differentiation,* Vol. 7, eds. Reinert, J. and Holtzer, H.,

(Springer Verlag, Berlin) pp.249-290.
4. Cohen, L.H., Newrock, K.M. and Zweidler, A. (1975) Science 190, 994-997.
5. Newrock, K.M., Cohen, L.H., Hendricks, M.B., Donnelly, R.J. and Weinberg, E.S. (1978) Cell 14, 327-336.
6. Brandt, W.F., Strickland, W.N., Strickland, M., Carlisle, L., Woods, D. and Von Holt, C. (1979) Eur.J.Biochem. 94, 1-10.
7. Birnstiel, M.L., Portmann, R., Busslinger, M., Schaffner, W., Probst, E. and Kressmann, A. (1978) in *Specific Eukaryotic Genes,* Alfred Benzon Symposium 13, eds. Engberg, J., Klenow, H. and Leick, V. (Munksgaard, Copenhagen), pp.117-129.
8. Schaffner, W., Gross, K., Telford, J. and Birnstiel, M. (1976) Cell 8, 471-478.
9. Gross, K., Schaffner, W., Telford, J. and Birnstiel, M. (1976) Cell 8, 479-484.
10. Portmann, R., Schaffner, W. and Birnstiel, M. (1976) Nature 264, 31-34.
11. Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H.O. and Birnstiel, M. (1978) Cell 14, 655-671.
12. Grosschedl, R. and Birnstiel, M.L. (1980) Proc.Natl.Acad.Sci.USA, in press.
13. Busslinger, M., Portmann, R. and Birnstiel, M.L. (1979) Nucleic Acids Res. 6, 2997-3008.
14. Hentschel, Ch., Irminger, J.C., Bucher, Ph. and Birnstiel, M.L. (1980) submitted to Nature.
15. Clarkson, S.G., Smith, H.O., Schaffner, W., Gross, K. and Birnstiel, M.L. (1976) Nucleic Acids Res. 3, 2617-2632.
16. Murray, K., Murray, N.E. and Brammar, W.J. (1975) FEBS Proceedings Tenth Paris Meeting 38, 193-207.
17. Birnstiel, M.L., Schaffner, W. and Smith, H.O. (1977) Nature 266, 603-607.
18. Maxam, A.M. and Gilbert, W. (1977) Proc.Natl.Acad.Sci.USA 74, 560-564.
19. Boseley, P., Moss, T., Mächler, M., Portmann, R. and Birnstiel, M.L. (1979) Cell 17, 19-31.
20. Loening, U.E. (1969) Biochem.J. 113, 131.
21. Wienand, U., Schwarz, Z. and Feix, G. (1979) FEBS Lett. 98, 319-323.
22. Smith, H.O. and Birnstiel, M.L. (1976) Nucleic Acids Res. 3, 2387-2398.
23. Southern, E.M. (1975) J.Mol.Biol. 98, 503-517
24. Wouters-Tyrou, D., Sautiere, P. and Biserte, G. (1976) FEBS Lett. 65, 225-228.
25. De Lange, R.J., Fambrough, D.M., Smith, E.L. and Bonner, J. (1969) J. Biol.Chem. 244, 319-334.
26. Ogawa, Y., Quagliarotti, G., Jordan, J., Taylor, C.W., Starbuck, W.C. and Busch, H. (1969) J.Biol.Chem. 244, 4387-4392.
27. De Lange, R.J., Hooper, J.A. and Smith, E.L. (1973) J.Biol.Chem. 248, 3261-3274.
28. Hooper, J.A., Smith, E.L., Sommer, K.R. and Chalkley, R. (1973) J.Biol. Chem. 248, 3275-3279.
29. Brandt, W.F., Strickland, W.N. and Von Holt, C. (1974b) FEBS Lett. 40, 407-429.
30. Brandt, W.F. and Von Holt, C. (1974a & b) Eur.J.Biochem. 46, 407-429.
31. Wouters-Tyrou, D., Sautiere, P. and Biserte, G. (1978) Eur.J.Biochem. 90, 231-239.
32. Yeoman, L.C., Olson, W.O., Sugano, N., Jordan, J.J., Taylor, C.W., Starbuck, W.C. and Busch, H. (1972) J.Biol.Chem. 247, 6018-6023.
33. Laine, B., Kmiecik, D. and Sautiere, P. (1978) Biochimie 70, 147-150.

34. Blankstein, L.A., Stollar, B.D., Franklin, S.G., Zweidler, A. and Levy, S.B. (1977) Biochemistry 16, 4557-4562.
35. Laine, B., Sautiere, P. and Biserte, G. (1976) Biochemistry 15, 1640-1645.
36. Durham, J.W. (1966) in *Treatise on Invertebrate Paleontology,* Part U, Echinodermata 3, ed. Moore, R.C. (University of Kansas Press, Lawrence, Kansas), pp. 431-433.
37. Brandt, W.F. and Von Holt, C. (1978) Biochim.Biophys.Acta 263, 351-367.
38. Strickland, M., Strickland, W.N., Brandt, W.F., Von Holt, C., Lehmann, A. and Wittmann-Liebhold, B. (1978a) Eur.J.Biochem. 89, 443-452.
39. Koostra, A. and Bailey, G.S. (1978) Biochemistry 17, 2504-2510.
40. Van Helden, P.D., Strickland, W.N., Brandt, W.F. and Von Holt, C. (1979) Eur.J.Biochem. 93, 71-78.
41. Isenberg, I. (1979) Ann.Rev.Biochem. 48, 159-191.
42. Iwai, K., Hayashi, H. and Ishikawa, K. (1972) J.Biochem.(Tokyo) 72, 357-367.
43. Ohe, Y., Hayashi, H. and Iwai, K. (1979) J.Biochem.(Tokyo) 85, 615-624.
44. Kedes, L.H., Cohn, R.H., Lowry, J.C., Chang, A.C.Y. and Cohen, S.N. (1975) Cell 6, 539-369.
45. Sures, I., Lowry, J. and Kedes, L.H. (1978) Cell 15, 1033-1044.
46. Grunstein, M. and Grunstein, J.E. (1977) Cold Spring Harbor Symp. Quant. Biol. 42, 1083-1092 and personal communication.
47. Overton, G.C. and Weinberg, E.S. (1978) Cell 14, 247-257.
48. Kunkel, N.S. and Weinberg, E.S. (1978) Cell 14, 313-326.
49. Childs, G., Maxson, R. and Kedes, L.H. (1979) Develop.Biol. 73, 153-173.
50. Fitch, W.M. and Margoliash, E. (1970) Evol.Biol. 4, 76.
51. Gannon, F., O'Hare, K., Perrin, F., Le Pennec, J.P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B. and Chambon, P. (1979) Nature 278, 428-434.
52. Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F. and Chambon, P. (1979) Nature 282, 567-574.
53. Nishioka, Y. and Leder, P. (1979) Cell 18, 875-882.
54. Konkel, D.A., Haizel, J.V. and Leder, P. (1979) Cell 18, 865-873.
55. Van Ooyen, A., Van den Berg, J., Mantei, N. and Weissmann, C. (1979) Science 206, 337-344.
56. Tsujimoto, Y. and Suzuki, Y. (1979) Cell 18, 591-600.
57. Smith, M., Leung, D.W., Gillan, S., Astell, C.R., Montgomery, D.L. and Hall, B.D. (1979) Cell 16, 753-761.
58. Tonegawa, S., Maxam, A.M., Tizard, R., Bernard, O. and Gilbert, W. (1978) Proc.Natl.Acad.Sci.USA 75, 1485-1489.
59. Ziff, E.B. and Evans, R.M. (1978) Cell 15, 1463-1475.
60. Surrey, S. and Nemer, M. (1976) Cell 9, 589-595.
61. Faust, M., Millward, S., Duchastel, A. and Fromson, D. (1976) Cell 9, 597-604.
62. Dickson, R.C., Abelson, J., Barnes, W.M. and Reznikoff, W.S. (1975) Science 187, 27-34.
63. Schaller, H., Gray, C. and Hermann, K. (1975) Proc.Natl.Acad.Sci.USA 72, 737-741.
64. Pribnow, D. (1975) Proc.Natl.Acad.Sci.USA 72, 784-788.
65. Ogata, R.T. and Gilbert, W. (1978) Proc.Natl.Acad.Sci.USA 75, 5851-5854.
66. Siebenlist, U. (1979) Nature 279, 651-652.