This is to certify that the

dissertation entitled

CHICKEN CHROMOSOMAL PROTEIN GENES

presented by

David Lawrence Browne

has been accepted towards fulfillment
of the requirements for

Ph. D. degree in Biochemistry

Major professor

Date July 20, 1990

CHICKEN CHROMOSOMAL PROTEIN GENES

By

David Lawrence Browne

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Biochemistry
1990

ABSTRACT

CHICKEN CHROMOSOMAL PROTEIN GENES

by

David Lawrence Browne

Two types of chicken chromosomal protein genes have been studied, the histone H1 genes and the HMG-14 genes. A known chicken H1 gene was used as a hybridization probe to isolate 4 other members of the gene family. The sequence of one of these genes was completely determined. Comparison of the two H1 sequences showed that the two genes encode unique proteins. The newly characterized gene contains common sequence elements in the 5' and 3' regions which flank the coding sequence. Partial sequence analysis showed that the other members of this gene family encode unique proteins and contain common eukaryotic promoter elements.

Chicken HMG-14 and HMG-17 cDNA clones were used to isolate the chicken genes for these proteins. Sequence analysis showed that the HMG-17 gene consists of 6 exons and 5 introns. The HMG-17 promoter contains the common

which is present in some, but not all, of the HMG-14 mRNA. The multiple splicing seen at the 5' end of the HMG-14 gene may result from multiple initiation sites, since the HMG-14 promoter contains no TATAA or CCAAT elements. Multiple processing also occurs at the 3' end of the HMG-14 transcript, in contrast to the HMG-17 transcript.

To Linda,

whose Spartan support

of science

has been

remarkable.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Literature Review: Histones and Histone Genes

In 1884 Albrecht Kossel reported the isolation of a fraction of basic proteins from the nuclei of goose erythrocytes (1). He called this fraction "histone" and discussed the possibility of its interaction with nucleic acids. Since his observations histones have been found to be essentially universal in eukaryotic nuclei, and much has been learned about the structures of histone proteins and the nature of their interactions with DNA. As our knowledge increases, though, so does the sophistication of our ignorance, and the biochemistry of histones remains an active field of study today.

Chromatin is the complex of proteins and nucleic acids found in the nuclei of all eukaryotic cells. It contains the genetic information of a cell and is intimately involved in both the expression of that information (transcription) and the maintenance of that information from one generation of cell to the next (replication). Chromatin undergoes dynamic structural changes as it performs each of these functions (2-4).

Eukaryotic chromosomes are composed of approximately equal amounts of DNA, histones, and other proteins. Histones are the main protein components of chromatin. They

histone is at least 10 times as abundant as any nonhistone chromatin protein (5).

Chromosomes have long been studied by microscopists, and biochemists have studied the components of chromatin. Around 1975 both types of studies revealed the basic repetitive nature of chromatin. Improved electron micrographs revealed the "beads on a string" structure of chromatin (6-8). This structure was correlated with the products of nuclease digestion of chromatin, which appeared to be monomers and multimers of a basic structure. This basic structure is the nucleosome.

When chromatin is digested with a nuclease like DNase I, nucleosomes are released. These particles contain about 200 base pairs (bp) of DNA and a standard complement of histones: one molecule of H1 and 2 molecules of each of the other classes of histones (9). Further digestion of nucleosomes releases H1 and reduces the DNA to 165 bp (10). The resulting nucleosome core is relatively resistant to further nuclease digestion.

The stability of the nucleosome cores is reflected by the ability to reconstitute them. Isolated core histones will reaggregate when mixed in solution, and when DNA is

165 bp of DNA are wrapped twice around the histone core. In undigested chromatin, cores are joined by a species-specific length of "linker" DNA (averaging about 30 bp) which is associated with H1 histone. One conception of this structure is shown in Figure 1.

The location of H1 external to the nucleosome core is inferred by its release by nucleases and by other studies of whole chromatin. H1 is more readily extracted from chromatin than are other histones; that is, lower concentrations of salt disrupt its interactions with DNA and other histones (17). Isolated H1 will reassociate with H1-depleted chromatin (18). The lower stability of H1 in chromatin has made the stoichiometry of H1 somewhat problematic. Measurements of the stoichiometry of H1 show there are about 0.8-1.0 molecules of H1 per core octamer (22). The absence of H1 from some nucleosomes may be artifactual or may reflect the in vivo situation.

The orientation and interactions of H1 in native chromatin are not known. It can be shown that H1 interacts with both ends of the 165 bp of core DNA (19) and crosslinking studies of reconstructed chromatin show that one H1 molecule can link two core particles (20,21).

Figure 1. Nucleosome structure. From (88).

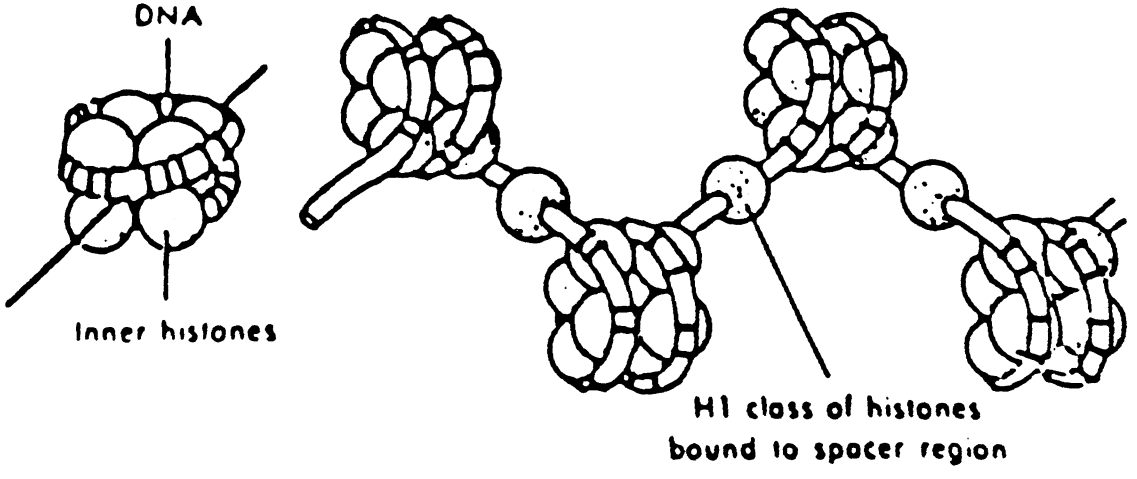

DNA
Inner histones
H1 class of histones
bound to spacer region

Figure 1

Electron microscopy of H1-depleted chromatin and reconstructed chromatin shows that H1 orders nucleosomes into a regular, zig-zag structure (23). Figure 2 illustrates this role of H1.

Histone structure

Because of their abundance amino acid sequences of histones from many species have been determined (24). When the structures of histones from many sources are compared some general features are seen. Histones are small globular proteins with molecular weights between 15 kilodaltons (kd) and 21 kd. They are rich in basic amino acids, with lysine and arginine typically comprising about 25% of the residues. These and other polar residues are concentrated in the amino-terminal regions of the molecule; the carboxy-terminal two-thirds is quite hydrophobic. H1 differs from the other histones by having another basic domain at the C-terminus. The amino-terminal basic domain and the hydrophobic central portion of H1 are approximately the same sizes as the equivalent domains of the core histones (25).

Sequence comparisons show that the primary structures of the histones have been highly conserved through time. The evolutionary stability of H4 is legendary. The sequences of H4 of calf, mouse, and frog are identical

Figure 2. The role of H1 in nucleosome ordering. From (89).

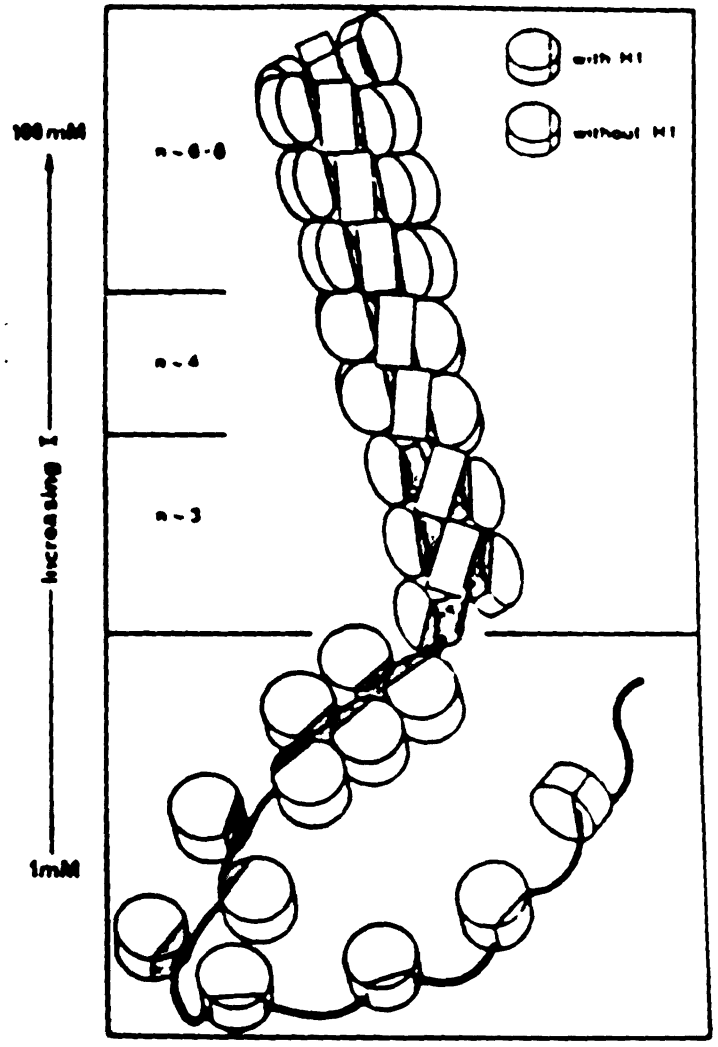


Figure 2

sequence (29), relative to that of vertebrates. H3 is also highly conserved. A calf H3 differs from a chicken H3 by four substitutions (30) and from pea H3 by only 5 substitutions (31).

H2A and H2B sequences have diverged more than H3 and H4 sequences. While mouse and calf H2B differ by only one residue, H2B from wheat has only 78% homology to the mammalian protein (32). Almost all of the variability in the H2B family is found in the charged amino-terminal domain, and includes deletions and insertions of residues as well as substitutions (33,34). The hydrophobic carboxy-terminal domain of H2B has been highly conserved. The H2A family has variability throughout the molecule. H2A proteins from chicken and calf are 78% homologous, and the proteins from plants (35) and yeast (36) are quite different from the vertebrate proteins.

The H1 class is the most divergent of the 5 classes of histones. In fact, H1 has never been found in yeast (37). Both ends of the molecule are variable and very basic; the central hydrophobic region is more conserved (38). H1 molecules vary much more in size than other histones: H1 proteins of 189 and 224 amino acids have been reported (39,40).

histone variants in individual organisms has long been known, and it originally led to a plethora of nomenclatures (41). There are two sources of histone variability within a species. First, non-allelic variants with different primary structures may exist. Second, all the histones undergo a variety of post-translational modifications.

The intraspecies evolutionary variability of different histone classes is similar to the variability seen among species. Thus, organisms generally have only one type of H4 histone protein. Three H3 histone proteins with different primary structures are known in calf, and they are identical to the 3 chicken types. H2A and H2B variants have diverged to the point that homologies between species can be difficult to assign (42,43).

The highly diverse H1 family includes some of the best studied variants. The most abundant of these are H1° and H5. H1° is a mammalian H1 variant which is found in high amounts in non-dividing tissues (44). H1° itself can be separated into two forms, but the differences between them are not known. The other major H1 variant is H5, which is found only in the nucleated erythrocyes of birds, reptiles, amphibians, and fish (45). Erythrocytes are non-dividing cells and, interestingly, H5 proteins are more like H1° than

number and amounts of H1 variants can vary from one tissue to another within a single species (48). In most cases primary structures of the variants have not been determined, so heterogeneity in this class of histones could result from post-translational modifications in addition to differences in primary structures.

All 5 classes of histones undergo a variety of post-translational modifications. The most common modifications are acetylation, methylation, and phosphorylation, but other modifications have also been found.

Two types of amino-acetylation are seen in histones: N-terminal and internal. The N-terminus of H1, H2A, and H4 can be acetylated (49-51). Serine is the usual acetylated N-terminal residue, but others are possible (52,53). Most histone acetylation occurs at the side chains of internal lysine residues of all the histones except H1. Internal acetylation is reversible. The general level of histone acetylation is controlled by the action of a variety of nuclear acetylases (54) and deacetylases (55,56). Since multiple side chains may be acetylated, and this modification is reversible, internal acetylation can generate many specific forms of histones (57).

Internal lysine side chains can also be methylated to

methylation of arginine side chains has also been found (58).

Phosphorylation of internal residues is a common modification that is seen in all 5 histone classes. Serine and threonine are the usual sites of phosphorylation, but 2 histidine residues of H4 can also be phosphorylated (59).

Many people have attempted to correlate histone variants and modifications with functional states of chromatin. Particularly popular are studies of cells undergoing differentiation, neoplastic transformation, senescence, and DNA repair. Despite hundreds of studies, the involvement of histone variants in chromatin activity is still poorly characterized and often controversial.

Studies of the roles of histone variants and modifications in modulating chromatin structure and function are difficult for two related, fundamental reasons. First, isolation of chromatin or fractions of chromatin must involve some destruction of structure. Second, specific variant structures are only a small fraction of total chromatin, and procedures to enrich chromatin preparations for specific variants are disruptive and have not been very successful. Only in a few special cases, involving tissue

changes that are part of this process. The best documented is the coupling of chromosome condensation and phosphorylation of specific residues of H1 (60). A mitosis-associated kinase, active during late G2 in the cell cycle, is responsible for phosphorylation of hydroxyl groups in the sequence lys-ser/thr-pro (61,62) in both the N-terminal and C-terminal domains of H1. The activity of this kinase is itself regulated by phosphorylation. It participates in a cascade of phosphorylation which leads to modifications of H1 histones, ribosomal proteins, and other nonchromosomal proteins (63). Other histone modifications occur at this time, including H3 phosphorylation (62) and an increase of poly-ADP-ribosylation of undetermined chromosomal proteins (64). These changes have been reported in other stages of the cell cycle, though (65-67), so their importance in mitosis is unclear.

The best biochemical studies of chromatin have allowed only vague correlations of structure and function. It seems that new approaches to the study of chromatin are needed. One that should prove fruitful is the increasingly refined determination by X-ray crystallography of the physical structure of nucleosomes of defined composition, including H1 and modified core histones.

a sea urchin ovum initiates a series of rapid cell divisions with no actual growth of the embryo. During this morula stage the need for histones and other nuclear proteins is great and at least 60% of the embryo's translational activity is devoted to the production of these proteins (68,69). Large numbers of synchronized early embryos are easy to obtain and are a good source of polysomes which are enriched with histone mRNA.

Measurements of the hybridization kinetics of sea urchin histone mRNA and genomic DNA showed that the histone genes occur several hundred times in the sea urchin genome (70). When the DNA was fractionated by equilibrium density gradient centrifugation most of these highly reiterated genes (if not all) were found in a satellite fraction (70,71). This suggested that the histone genes might be linked, but proof of this was only obtained after the development of recombinant DNA techniques when cloned genomic DNA became available.

When histone mRNAs were used to probe cloned DNA and restricted genomic DNA it was learned that genes of the 5 classes of histones are grouped into nearly identical clusters, each about 6 kilobase pairs (Kb) long, which are tandemly repeated 300-600 times. Figure 3a shows an example

strand of DNA. The order of the genes (...H1-H4-H2B-H3-H2A-
...) has been conserved. The sequences between the coding regions are A+T-rich and contain regulatory elements (72) but not other genes.

The availability of cloned histone genes allowed study of histone gene organization in other organisms. When the organization of Drosophila histone genes was investigated, it was found that these genes, too, occur as tandemly reiterated quintets. The Drosophila repeats differ from sea urchin repeats in several ways, though. The order of the histone genes in these flies is different (...H1-H3-H4-H2A-H2B-...) and the genes are not all transcribed from the same strand of DNA. Also, 2 types of repeats are found. They differ by the presence or absence of a specific 240 bp intergenic sequence (73). This insert doesn't seem to be important for histone production, since flies have been bred which lack either type of repeat with no apparent deleterious effects (74,75). Figure 3b shows the structure of the Drosophila histone repeats.

The arrangement of histone genes is less ordered in the vertebrates. Reiterated quintets of histone genes occur in amphibians, but a good deal of heterogeneity has evolved

clusters of histone genes (77). The number of copies of each class of histone gene ranges from about 50 to 1500 in amphibians, probably because of the large range of haploid DNA content among these species (78).

Birds and mammals have fewer histone genes than those animals discussed above, about 10-40 copies of each type in a haploid genome. These genes are not arranged in repeated blocks at all. Many of the genes are clustered, though. Figure 3c shows restriction maps of a number of clones of chicken genomic DNA which contain histone genes, as judged by Southern blotting. It is clear that, though these clones contain clusters of histone genes, each is unique and not tandemly repeated.

The collection of genomic clones shown in Figure 3c contains most, but probably not all, of the chicken histone genes. Analysis of genomic Southern blots suggests that the H1 gene is present in about 6 copies, while the other histone genes are present about 10-15 times each (79). These estimates agree with early studies of solution hybridization kinetics (80) which were used to infer histone gene copy number.

The sequence of at least one gene of each class of chicken histone has been determined (81). Also, various sea

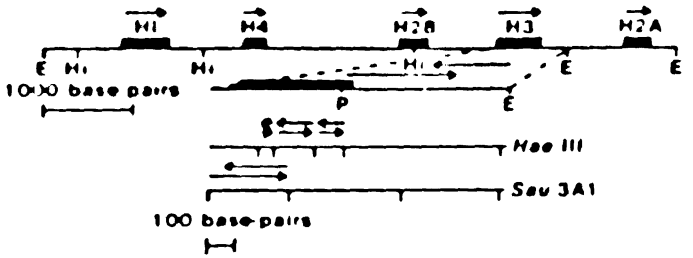
Figure 3. The genomic organization of histone genes.

a.) The major repeat of sea urchin L. pictus. From (87).

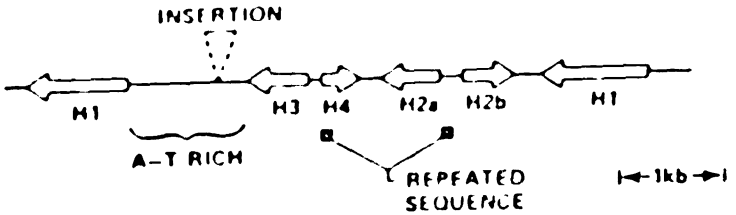
b.) The Drosophila melanogaster repeat. From (73).

c.) Chicken histone genes. Location of genes, as determined by Southern blot analysis, is indicated by boxes over the restriction maps. Solid black = H1, all white = H2A, vertically hatched = H2B, diagonally hatched = H3, and stippled = H4. Sequenced genes are named and their direction of transcription is indicated with arrows. From (81).

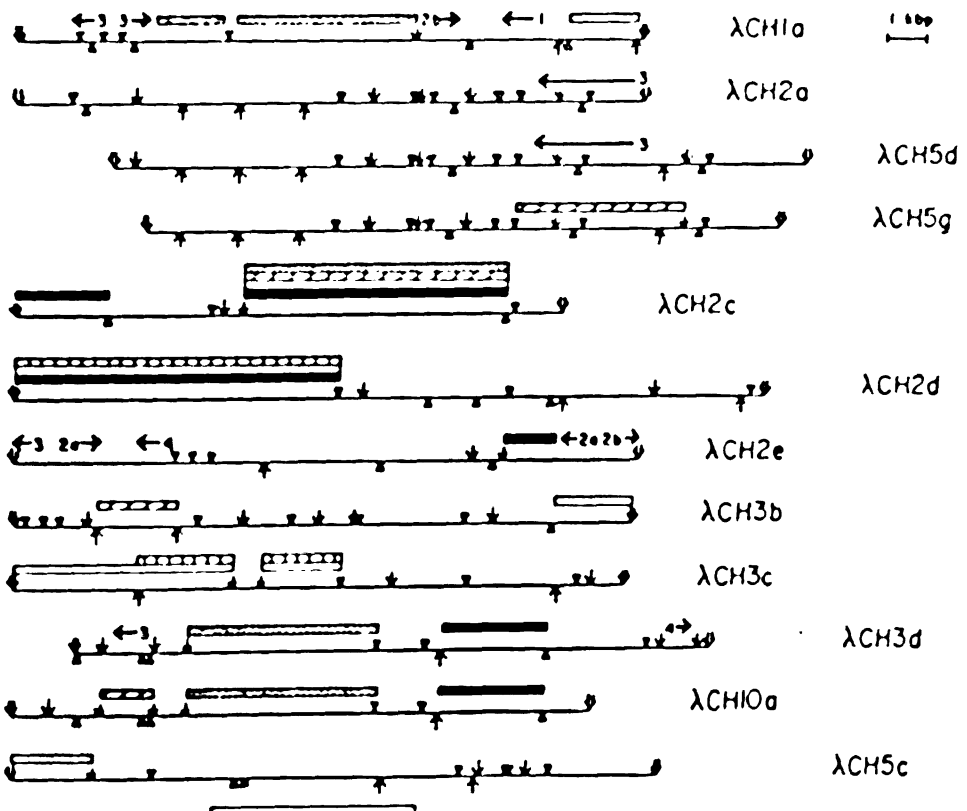
A.


H1
H4
H2B
H3
H2A
E
H1
H1
H1
E
E
1000 base pairs
P
E
Hae III
Sau 3A1
100 base pairs

B.


INSERTION
H1
H3
H4
H2a
H2b
H1
A-T RICH
REPEATED
SEQUENCE
1kb

C.


λCH1a
λCH2a
λCH5d
λCH5g
λCH2c
λCH2d
λCH2e
λCH3b
λCH3c
λCH3d
λCH10a
λCH5c

The regions 5' of the coding sequences of genes contain sequences which are important for gene expression. The best characterized of these promoter elements are the TATAA element (82) and the CCAAT element (83). The TATAA element has been found upstream of almost every histone gene analyzed, as well as many other genes. This sequence directs the start of transcription to a point about 30 bp downstream of it through its interaction with RNA polymerase II and other proteins (83). Some genes lack the TATAA element, and these genes often initiate transcription at several sites. The CCAAT element is also found in many genes, including most histone genes examined to date. Its location, 30-60 bp upstream of the TATAA element, is variable, and some genes contain more than one copy. The CCAAT element, like the TATAA element, interacts with proteins to modulate transcription.

Other upstream sequences may be promoter elements which are specific to one gene or a family of genes. The functions of these elements are often unknown, but they are probably binding sites for less ubiquitous activators of transcription. Comparisons of vertebrate histone genes have revealed several sequences that are evolutionarily conserved and thus are candidate promoter elements. Two of these will

representatives of a given class in a given species and often in the same class of histone genes in wide variety of species.

Some histone H2B genes are preceded by the octamer sequence ATTTGCAT. This element may be involved in cell cycle regulation of the H2B genes by interaction with one or more regulatory proteins (84). The octamer sequence has been shown to bind a ubiquitous transcription factor Oct1. This sequence is also present in lymphoid-specific enhancers and can bind a lymphoid-specific factor, Oct2.

A number of histone H1 genes have an A+C-rich element located upstream of the CCAAT element (85). The function of this element (usually AAACACA) is not known but its evolutionary conservation suggests it is important in the regulation of these genes.

DNA sequencing of histone coding regions has revealed that almost all histone genes consist of a single exon. Two of the few exceptions occur in chickens (86). The intron-containing chicken histone genes are also unusual in that they are unlinked to other histone genes, and their expression does not vary through the cell cycle.

Histone mRNA differs from most other mRNA because it is generally not polyadenylated. Histone mRNA usually ends

stability on histone mRNA. This feature of histone mRNA has been highly conserved and is found in animals from sea urchins to mammals.

Proposal

Determination of the structure of genes is useful for two main reasons. First, comparisons of gene structures have allowed identification of sequences important for the functions of both the genes and their products. Second, knowledge of a gene's structure allows very sensitive measurements of its activity.

Characterization of the H1 family is the only definitive way to determine variability in this important class of proteins, and to differentiate variability in primary sequence from post-translational variability. Only then can the participation of such variability in different states of chromatin structure and, perhaps, in chromatin function (transcription/replication) be assessed. H1 histone gene structure analysis will also facilitate further studies of H1 function by enabling the measurement of the production of specific H1 variants at the transcriptional level. The high level of similarity in the coding regions of the 6 chicken histone genes requires that individual DNA sequences be determined in order to develop gene-specific

We propose to:

1. Determine the size of the H1 family in chickens by isolating all of the genes.
2. Determine the sequences of the genes to deduce primary sequence variability and sequences important for gene activity.
3. Use knowledge of H1 gene structure to measure the activity of the genes in chicken cells.

References

1. Donecke, D., and P. Karlson. (1984). Trends in Biochemical Sciences 9:404-406.

2. Weintraub, H., and M. Groudine. (1976). Science 93:848.

3. Weisbrod, S. (1982). Nature (London) 297:289.

4. Felsenfeld, G. (1978). Nature (London) 271:115.

5. Johns, E.W.,(ed.). (1983). "The HMG Chromosomal Proteins." Academic Press, New York.

6. Woodcock, C.L.F. (1973). J. Cell. Biol. 59:368.

7. Olins, A.L., and D.E. Olins. (1973). J. Cell Biol. 59:252.

8. Olins, A.L., and D.E. Olins. (1974). Science 183:330.

9. Kornberg, R.D. (1974). Science 184:868.

10. Noll, M., and R.D. Kornberg. (1977). J. Mol. Biol. 109:393.

11. Thomas, J.O., and R.D. Kornberg. (1975). Proc. Nat. Acad. Sci. USA 72:2626.

12. Stein, A., M. Bina-Stein, and R.T. Simpson. (1977). Proc. Nat. Acad. Sci. USA 74:2780.

13. Thomas, J.O., and R.D. Kornberg. (1975). FEBS Lett. 58:353.

14. Thomas, J.O., and P.J.G. Butler. (1977). J. Mol.

16. Thomas, J.O., and P.J.G. Butler. (1978). Cold Spring Harbor Symp. Quant. Biol. 42:119.

17. Steinmetz, M., R.E. Streeck, and H.G. Zachau. (1978). Eur. J. Biochem. 83:615.

18. Caron, F., and J.O. Thomas. (1981). J. Mol. Biol. 146:513.

19. Finch, J.T., L.C. Lutter, D. Rhodes, T. Richmond, B. Rushton, M. Levitt, and A. Klug. (1977). Nature (London) 269:29.

20. Itkes, A.V., B.O. Glotov, L.G. Nikolaiv, S.R. Preem, and E.S. Severin. (1980). Nuc. Acids Res. 8:503.

21. Thomas, J.O., and A.J.A. Khabaza. (1980). Eur. J. Bioch. 112:501.

22. Bates, D.L., and J.O. Thomas. (1981). Nuc. Acids Res. 9:5883.

23. Thoma, F., T. Koller, and A. Klug. (1979). J. Cell. Biol. 83:403.

24. Isenberg, I. (1979). Ann. Rev. Bioch. 48:159.

25. Cole, R.D. (1984). Anal. Bioch. 136:24.

26. Seiler-Tuyns, A., and M.L. Birnsteil. (1981). J. Mol. Biol. 151:607.

27. Moorman, A.F.M., P.A.J. de Boer, R.J.M. deLaaf, W.M.A.M. Van Dongen, and O.H.J. Destree. (1981). FEBS

29. DeLange, R.J., D.M. Fambrough, E.L. Smith, and J. Bonner. (1969). J. Biol. Chem. 244:5669.

30. Brandt, W.F., and C. Von Holt. (1974). Eur. J. Bioch. 46:407.

31. Patthy, L., E.L. Smith, and J. Johnson. (1973). J. Biol. Chem. 248:6834.

32. Von Holt, C., W.N. Strickland, W.F. Brandt, and M.S. Strickland. (1979). FEBS Lett. 100:201.

33. Franklin, S.G., and A. Zweidler. (1977). Nature (London) 266:273.

34. Hunt, L.T., and M.O. Dayhoff. (1982). in "Macromolecular Sequences in Systematic and Evolutionary Biology", Goodman, M., (ed.), 193. Plenum Press, New York.

35. Rodrigues, J. de A., W.F. Brandt, and C. Von Holt. (1979). Biochim. Biophys. Acta 578:196.

36. Choe, J., P. Kolorubetz, and M. Grunstein. (1982). Proc. Nat. Acad. Sci. USA 79:1484.

37. Brandt, W.F., K. Patterson, and C. Von Holt. (1980). Eur. J. Biochem. 110:67.

38. Smith, B.J., M.R. Harris, C.M. Sigournay, E.L.V. Mayes, and M. Bustin. (1984). Eur. J. Biochem. 138:309.

39. Smith, B.J., J.M. Walker, and E.W. Johns. (1980).

41. Bradbury, E.M. (1975). in "The Structure and Function of Chromatin", Ciba Foundation Symposium 28. Ass. Scientific Publ., Amsterdam.

42. West, M.H.P., and W.M. Bonner. (1983). Comp. Biochem. Physiol. 76B:455.

43. Zweidler, A. (1984). in "Histone Genes: Structure, Organization, and Regulation", Stein, G.S., Stein, J.L., and Marzluff, W.F., (eds.), p. 339. Wiley/Interscience, New York.

44. Smith, B.J., and E.W. Johns. (1980). FEBS Lett. 110:25.

45. Isenberg, I. (1979). Ann. Rev. Biochem. 48:159.

46. Smith, B.J., J.M. Walker, and E.W. Johns. (1980). FEBS Lett. 112:42.

47. Smith, B.J., J.M. Walker, C.M. Sigournay, E.L.V. Mayes, and M. Bustin. (1984). Eur. J. Biochem. 138:309.

48. Cole, R.D. (1984). Anal. Biochem. 136:24.

49. Rall, S.C., and R.D. Cole. (1971). J. Biol. Chem. 246:7175.

50. Phillips, D.M.P. (1963). Biochem. J. 87:258.

51. Ogawa, Y., G. Quagliarotti, J. Jordan, C.W. Taylor, W.C. Starbuck, and H. Busch. (1969). J. Biol Chem. 244:4387.

53. Harvey, R.P., J.A. Whiting, L.S. Coles, P.A. Krieg, and J.R.E. Wells. (1983). Proc. Nat. Acad. Sci. USA 80:2819.

54. Saxholm, H.J.K., A. Pestana, L. O'Connor, D.A. Sattler, and H.C. Pitot. (1982). Mol. Cell. Biochem. 46:129.

55. Kelner, D.N., and K.S. McCarthy, Sr. (1984). J. Biol. Chem. 259:3413.

56. Yikioka, M., S. Sasaki, S. Le Qui, and A. Inoue. (1984). J. Biol. Chem. 259:8372.

57. Liao, L.W., and R.D. Cole. (1981). J. Biol. Chem. 256:3024.

58. Gupta, A., D. Jensen, S. Kim, and W.K. Paik. (1972). J. Biol. Chem. 257:9677.

59. Matthews, H.R., and V.D. Heubner. (1984). Mol. Cell. Biochem. 59:81.

60. Bradbury, E.M., R.J. Inglis, H.R. Matthews, and H. Sarner. (1973). Eur. J. Biochem. 33:131.

61. Quirin-Stricker, C. (1984). Eur. J. Biochem. 142:317.

62. Gurley, L.R., R.A. Walters, and R.A. Tobey. (1974). J. Cell Biol. 60:356.

63. Maller, J.L. (1990). Biochemistry 29:3157.

64. Tanuma, S., and Y. Kanai. (1982). J. Biol. Chem.

66. Whitlock, J.P., D. Goleazzi, and H. Schulman. (1983). J. Biol. Chem. 258:1299.

67. Kidwell, W.R., and M.G. Mage. (1976). Biochemistry 15:1213.

68. Kedes, L., and P. Gross. (1969). J. Mol. Biol. 42:559.

69. Moav, B., and M. Nemer. (1971). Biochemistry 10:881.

70. Kedes, L., and M. L. Birnstiel. (1971). Nature New Biology 230:165.

71. Birnstiel, M.L., J. Telford, E.S. Weinberg, and D. Strafford. (1974). Proc. Nat. Acad. Sci. USA 71:2900.

72. Hentschel, C., and M.L. Birnstiel. (1981). Cell 25:301.

73. Lifton, R.P., M.L. Goldberg, R.W. Karp, and D.H. Hogness. (1977). Cold Spring Harbor Symp. Quant. Biol. 42:1047.

74. Moore, G., J. Proncunier, D. Cross, and A. Grigliatti. (1979). Nature (London) 282:312.

75. Wright, T., R. Hodgetts, and A. Sherald. (1976). Genetics 84:267.

76. Stephenson et al. (1981). Cell 27:639.

77. Turner, Woodland. (1983). Nuc. Acids Res. 11:971.

78. Old, R.W., and H.R. Woodland. (1984). Cell 38:624.

80. Crawford, R.J., P. Krieg, R.P. Jarvey, D.A. Gewish, and J.R.E. Wells. (1979). Nature (London) 279:132.

81. Sugarman, B., J.B. Dodgson, and J. Engel. (1983). J. Biol. Chem. 258:9005.

82. Goldberg, M. (1979). Thesis. Standford University.

83. McKnight, S.L., and R. Kingsbury. (1982). Science 217:316.

84. Fletcher, C., N. Heintz, and R.G. Roeder. (1987). Cell 51:773.

85. Coles, L.S., and J.R.E. Wells. (1985). Nuc. Acids Res. 13:585.

86. Engel, J.D., B.J. Sugarman, and J.B. Dodgson. (1982). Nature (London) 297:434.

87. Roberts, S.B., K.E. Weisser, and G. Childs. (1984). J. Mol. Biol. 174:647.

88. Olins, D.E., and A. L. Olins. (1978). Amer. Sci. 66:704.

89. Thoma, F., T. Koller, and A. Klug. (1979). J. Cell Biol. 83:403.

Chapter 2

Materials and Methods

Materials

Restriction enzymes, calf alkaline phosphatase, T4 polynucleotide kinase, T4 DNA ligase, RNase A, S1 nuclease, DNase I, DNA polymerase I, and reverse transcriptase were obtained from the following sources: Bethesda Research Laboratories, United States Biochemical Corp., International Biotechnologies, Inc., Promega Biotec, New England Biolabs, and Boehringer Mannheim. Yeast tRNA was obtained from Sigma Chemical Company. γ-^{32}P-ATP and α-^{32}P-dCTP were obtained from New England Nuclear. The plasmid cloning vector "Bluescribe" and its *E. coli* host XL1-Blue were obtained from Stratagene. Oligonucleotide primers JD-20 (d[AATAGGAAGGGAACCGCCGAG]) and JD-21 (d[TTTTCGAGCGTCCT GAGGAA]) were obtained from the MSU Macromolecular Synthesis Facility. λgt11 primers were purchased from New England Biolabs. The primer "SK" for sequencing DNA cloned in Bluescribe was obtained from Stratagene. Fifteen day old chicken embryos were obtained from the MSU poultry farm.

Methods

The cloning and Southern blotting analysis used in this work followed the protocols outlined by Maniatis, Fritsch, and Sambrook (1). The H1 histone probe used to analyze native or cloned chicken DNA was a nick-translated PstI-

probes used to analyze chicken genomic clones were the nick-translated inserts of cDNA clones pLM3b (HMG-14) and pLG1a (HMG-17), which are described in the Appendix. The nick-translation protocol (1) routinely produced probes of 10^8 dpm/μg.

Isolation of HMG Clones

The library of recombinant λCharon 4A chicken genomic clones originally described by Dodgson, Strommer, and Engel (3) was screened with the HMG-14 and HMG-17 probes described above. 4×10^4 plaque forming units were plated on each of 10 150mm petri dishes for each screen, and duplicate filters were prepared by standard methods (1). Plaques which were positive in the initial screen were isolated, replated, and rescreened. Three genomic clones of each HMG gene were purified and analyzed by restriction mapping and sequence determination.

Sequence Determination

Chicken H1 histone, HMG-14, and HMG-17 gene sequences were determined by the chemical degradation method of Maxam and Gilbert (4) as modified by Smith and Calvo (5). The sequences of the 5' and 3' untranslated regions of the HMG-14 cDNA clone pLM2a were determined by the chain termination method of Sanger, Nicklen, and Coulson (6) using the

RNA Isolation

Total RNA was isolated from 15 day old chicken embryo liver, brain, heart, skeletal muscle, and blood by the method of Chirgwin et al. (7) as described by Davis, Dibner, and Batten (8). Approximately 40 embryos yielded 1-4g of each tissue. Tissues were homogenized in 16 ml of 4M guanidine isothiocyanate, 25mM sodium acetate (pH 6), 50mM 2-mercaptoethanol. The homogenate was layered onto 16 ml of 5.7M CsCl in an ultra-centrifuge tube and spun overnight (37,000 rpm, 20°C) in a Beckman Ti50 rotor. The clear, gelatinous RNA pellet was dissolved in 0.3M sodium acetate (pH 6) and extracted with phenol:chloroform (1:1). The RNA was precipitated from the aqueous phase with 2 volumes of cold ethanol, redissolved in 10mM Tris, 1mM EDTA (pH 7.6) and quantified spectrophotometrically at 260 nm.

S1 Protection Analysis

HMG-14 and HMG-17 mRNA was analyzed by S1 protection as described by Maniatis, Fritsch, and Sambrook (1). The HMG-17 probe was made by digesting 2μg of pHM1.7BH with Sau96a1, treating the digested DNA with calf alkaline phosphatase and the T4 polynucleotide kinase and γ-^{32}P-ATP. The 118 bp Sau96a1 fragment which contains the HMG-17 transcription

isolation of the labelled HincII-OxaNI fragment (approximately 700 bp) from a 1.2% agarose gel. DNA was labelled to a specific activity greater than $5x10^7$ dpm/μg by this method. Enough DNA to give $5x10^5$ dpm was coprecipitated with 50μg of RNA from embryonic tissue (or yeast tRNA) for use in the hybridization and digestion steps of the protocol. Hybrids were formed overnight at 50° in 30μl of 40 mM PIPES (pH6.4), 1mM EDTA, 0.4M NaCl, 80% formamide. After hybridization 300 μl ice-cold nuclease buffer (0.28M NaCl, 0.05M NaOAc, 4.5mM $ZnSO_4$) was added and the mixture divided into 2 equal aliquots. S1 nuclease (150u,300u) was added to each aliquot; digestion was performed for 30 minutes at 37°C. Products of the reactions were analyzed on autoradiographs of 6% denaturing gels (4).

Primer Extension

HMG-14 mRNA was analyzed by the primer extension method of McKnight and Kingsbury (9) as described by Ausubel (10). 100 ng of the oligonucleotide primer JD-21 was labelled with γ-^{32}P-ATP and polynucleotide kinase. 1 ng of labelled primer was coprecipitated with 50μg RNA, then hybridized to complementary sequences by overnight incubation in 1M NaCl, 165mM HEPES (pH 7.5), 0.33mM EDTA. After ethanol precipitation of the mixture, hybridized primer was extended

References

1. Maniatis, T., E. Fritsch, and J. Sambrook. (1982). "Molecular Cloning: A Laboratory Handbook." Cold Spring Harbor Laboratory, New York.

2. Sugarman, B., J.B. Dodgson, and J. Engel. (1983). J. Biol. Chem. 258:9005.

3. Dodgson, J.B., J. Strommer, and J.D. Engel. (1979). Cell 17:879.

4. Maxam, A., and W. Gilbert. (1980). Methods in Enzymology 65:499.

5. Smith, D., and J. Calvo. (1980). Nuc. Acids Res. 8:2255.

6. Sanger, F., S. Nicklen, and A.R. Coulson. (1977). Proc. Nat. Acad. Sci. USA 74:5463.

7. Chirgwin, J.M., A.E. Przybyla, R.J. MacDonald, and W.J. Rutter. (1979). Biochemistry 18:5294.

8. Davis, L.G., M.D. Dibner, and J.F. Batten. (1986). "Basic Methods in Molecular Biology", p. 130. Elsevier Sci. Pub. Co., New York.

9. McKnight, S.L., and R. Kingsbury. (1982). Science 217:316.

10. Ausubel, F.M., R. Brent, R.E. Kingston, D.D. Moore, J.G. Seideman, J.A. Smith, and K. Struhl. (1987).

Chapter 3

Sequence Analysis of Chicken H1 Histone Genes

Sugarman et al. (1) reported the isolation of 50 recombinant λCharon4A clones which contain chicken histone genes. Fifteen unique members of this collection were further characterized by restriction mapping and Southern blotting. They located an H1 gene on λCH1a and determined its sequence. When a fragment from this gene was used as a hybridization probe, 6 of the other clones were shown to contain H1 genes (Figure 4). Only two of these clones, λCH3d and λCH10a, were known to overlap.

To facilitate more detailed mapping of the H1 genes, plasmid subclones of fragments containing the genes were constructed. After DNA was isolated from each recombinant phage clone, it was digested with the appropriate restriction enzymes. The H1 gene-bearing fragments were purified from agarose gels and ligated into the common cloning vector pBR322. These subclones are indicated and named in Figure 4.

H1.10a

Detailed restriction maps of p2c2.2RK and p10a5.0BR were deduced from the fragment generated by single and double restriction digests (Figure 5). The known chicken H1 gene sequence (of H1.1a) contains overlapping ApaI and SacII sites at codons 36-39. A similar coincidence of ApaI and

Figure 4. Recombinant λ Charon 4A chicken genomic clones which contain chicken H1 genes, and plasmid subclones derived from them. From (1).

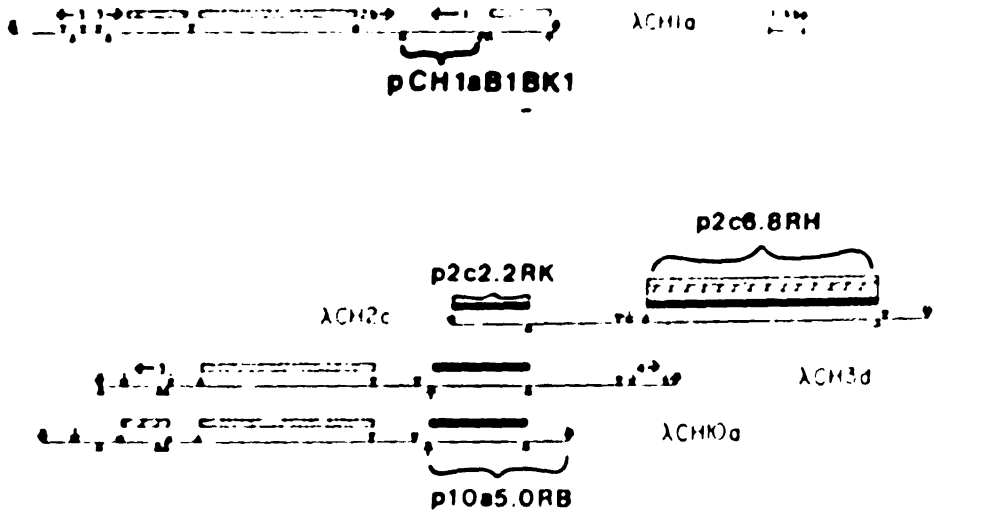

λCH1a
pCH1aB1BK1
p2c6.8RH
p2c2.2RK
λCH2c
λCH3d
λCHKOa
p10a5.0RB
p5e5.2RB
λCH5e
λCH2e
p2e3.5RR
p2d7.9RH
λCH2d

Figure 5. Comparison of p2c2.2RK and p10a5.0BR.

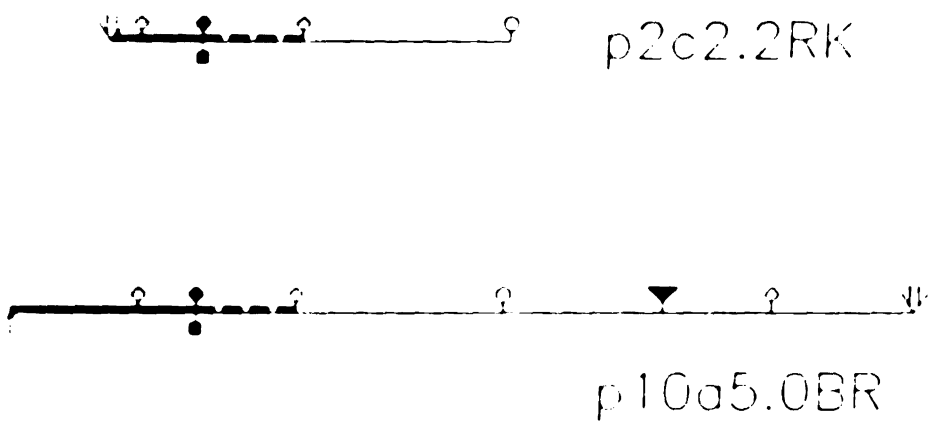

p2c2.2RK
p10a5.0BR
1 kb
EcoRI linker
BamHI
BstEII
KpnI
AvaI
ApaI
SacII

Figure 5

locate and orient the H1 genes on these two subclones. The strongly hybridizing region of each subclone contains most of an H1 gene (approximately 185 codons); the weakly hybridizing fragments each contain approximately 36 codons of sequence homologous to the 5' end of the H1.1a gene.

The conservation of restriction enzyme sites in the suspected coding regions could be expected because of the generally high homology among histone genes. This level of homology does not extend into the flanking regions of the histone genes. Grandy's analysis of the chicken H2B genes (2) showed that, except for a few short sequence blocks, sequences flanking the histone genes are not conserved. Since the maps of p2c2.2RK and p10a5.0BR are identical for more than 1.5 kb upstream of their H1 genes, these clones must overlap. The λclones from which they are derived are aligned in Figure 4 to show how they overlap.

The only discrepancy between the overlapping maps of λCH2c and λCH3d is in the location of the sequenced H4 gene on λCH3d. It seemed likely that the order of the small EcoRI fragments at the end of the λCH3d insert was incorrectly assigned. If these fragments were misordered, the map of the H4 gene-bearing fragment (called pCH3dR8 in (1)) should correspond to one end of p2c6.8RH. Figure 6

Figure 6. Restriction maps of pCH3dR8 and p2c6.8RH. Weak hybridization to the H1 probe is indicated by the dashed overline. Arrows indicate regions of p2c6.8RH which were sequenced to confirm the absence of H1 gene in this region.

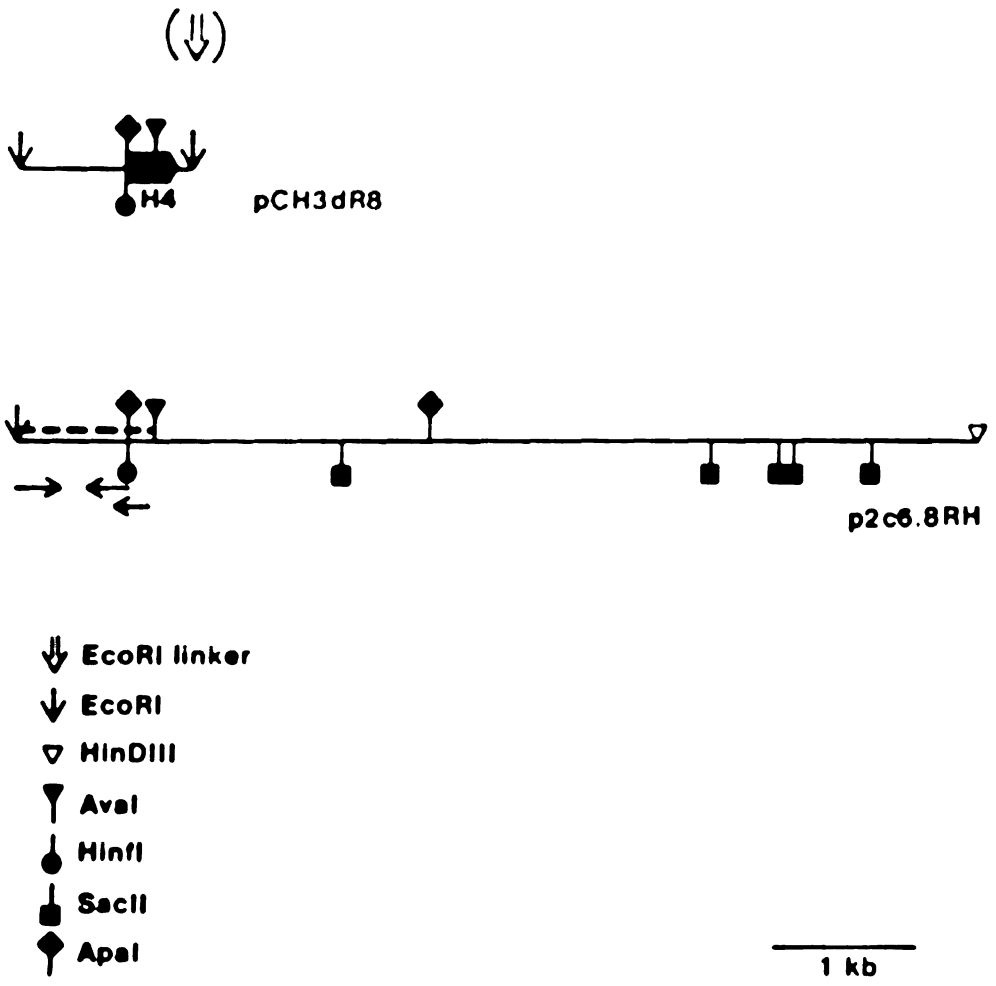

H4
pCH3dR8
p2c6.8RH
EcoRI linker
EcoRI
HinDIII
AvaI
HinfI
SacII
ApaI
1 kb

Figure 6

The region subcloned into p2c6.8RH was previously reported to hybridize to an H1 probe (1); however, we only observed weak hybridization localized near the H4 histone gene at its left-hand end (Figure 6). Furthermore, there is barely enough room between the H4 gene and the end of p2c6.8RH to contain a normal H1 gene. These facts cast doubt on the assignment of an H1 gene to p2c6.8RH. Sequence analysis upstream of the known H4 gene showed that there is in fact no H1 gene in this area (Figure 6, sequence not shown). To prove that there is no H1 gene on p2c6.8RH, restriction digests of chicken genomic DNA were blotted onto nylon membranes and probed with either the p2c6.8RH insert or the H1 gene probe. These probes hybridize to different genomic restriction fragments (data not shown), confirming that there is no H1 gene on p2c6.8RH.

H1.2e

Fine structure maps of p2e3.5RR and p5e5.2RB are shown in Figure 7. The patterns of hybridization to the H1 gene probe around the coincident ApaI and SacII sites locate and orient the H1 genes on these subclones.

The maps of these subclones are identical for more than 2 kb downstream of the H1 genes, so we concluded that these clones overlap. The parent λclones, λCH2e and λCH5e, are

Figure 7. **Restriction maps of p2e3.5RR and p5e5.2RB. Hybridization to the H1 gene probe is indicated by heavy or dashed overlining.**

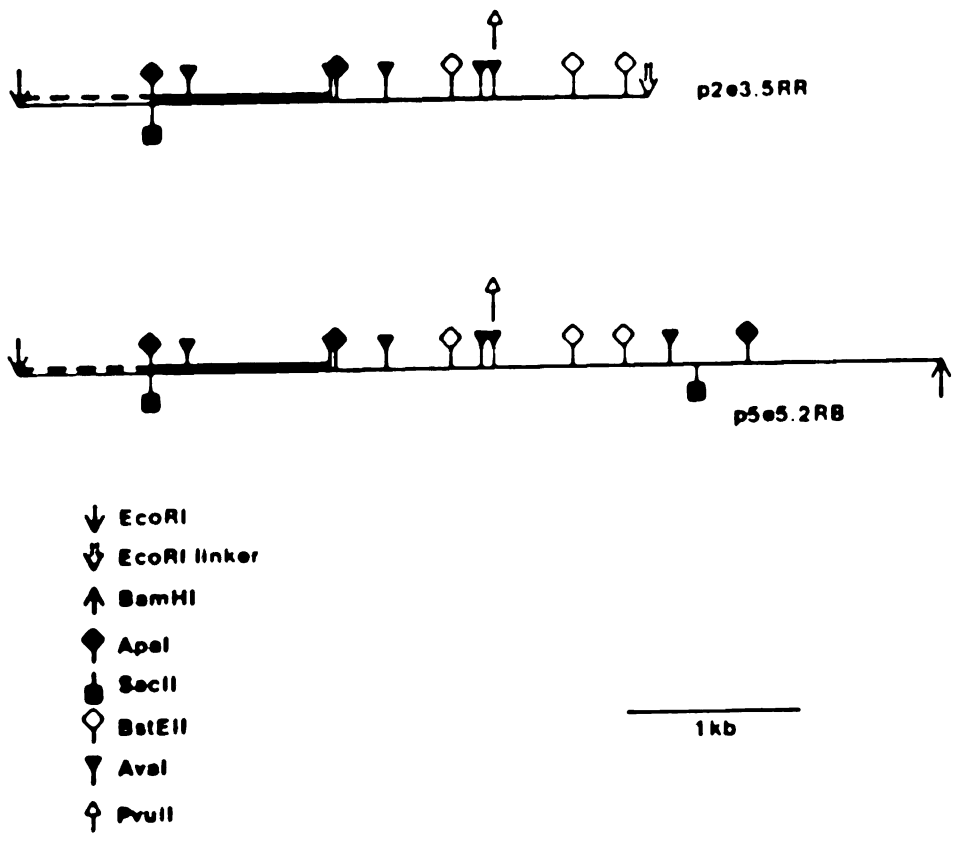

p2e3.5RR
p5e5.2RB
EcoRI
EcoRI linker
BamHI
ApaI
SacII
BstEII
AvaI
PvuII
1kb

Figure 7

The failure to assign an H2B gene to the region of p5e5.2RB was an oversight in the initial characterization of λCH5e (1).

H1.1c

An analysis of chicken histone gene copy number by Ruiz-Carillo et al. (3) showed that the chicken genome contains 5 or 6 H1 genes. Figure 4 shows that since λCH2c and λCH10a contain the same H1 gene, λCH2e and λCH5e contain the same H1 gene, and p2c6.8RH contains no H1 gene, the set of recombinant λ clones characterized by Sugarman et al. (1) contains 4 H1 genes.

Sugarman et al. had isolated 50 histone gene-bearing λrecombinants in their original library screen but only characterized 25 of them. We decided to screen the uncharacterized isolates for H1 genes in an attempt to find other members of the gene family. This was done by spotting each of the isolates in an array on a lawn of host bacteria. After overnight incubation each clone formed a large (8 mm) plaque. The arrays were transferred to nylon membranes and hybridized to the H1 gene probe.

Four uncharacterized clones showed homology to the probe. Restriction mapping showed that 3 of these were identical to previously characterized clones. One clone was a new type. The map of this clone, λCH1c, is shown in

Figure 8. Restriction map of λCH1c. The region which hybridizes to the H1 probe is indicated by the heavy bar. The plasmid subclone of this region is p1c2.0HH.

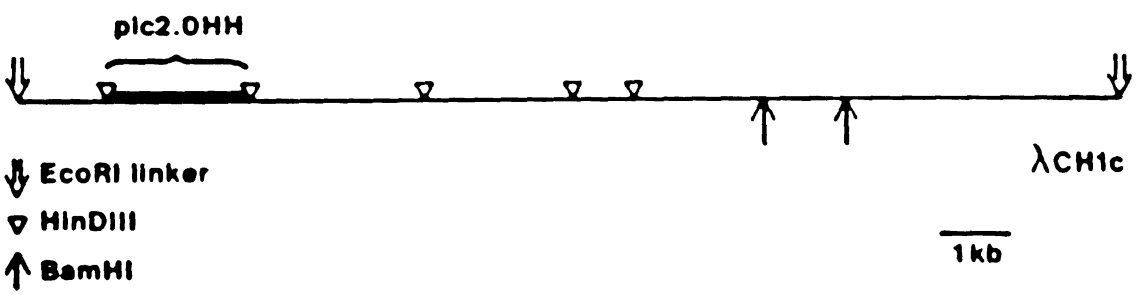


Figure 8

family to be isolated. Since the original set of 50 λ recombinants contain numerous sibling pairs and overlapping clones, and since no more than 5 genomic bands were found to hybridize to a H1 gene probe by Ruiz-Carillo et al., it seemed that the H1 gene family was complete with 5 members.

This idea was wrong. D'Andrea et al. were studying the organization of the chicken histone genes using the same recombinant λCharon4A chicken clones as we were (18). Their collection of λ clones, like ours, contains 5 H1 histone genes. They also isolated a cosmid (cosmid 6.3c) which contains a sixth chicken H1 histone gene. This gene lies very near sequences contained in λCH2d, and lies on the same genomic EcoRI, HinDIII, and BamHI fragments as H1.2d. These were the enzymes we used to determine H1 histone gene copy number by Southern blotting, which explains our failure to detect a sixth gene. We attempted to isolate this gene from our library, using probes D'Andrea et al. showed flank it. We were unsuccessful, so it seems this part of the chicken genome is not represented in the library.

H1 sequences

Sugarman et al. published the first DNA sequence of a chicken H1 gene (1). To learn the degree of diversity among members of the H1 family: 1) the entire coding sequence of the H1 gene on λCH2d (H1.2d) was determined, and also 5'

H1.10a sequence

The sequence of H1.10a was determined by the chemical degradation technique of Maxam and Gilbert. Figure 9 shows the restriction sites which were radioactively labeled with ^{32}P-ATP and polynucleotide kinase to sequence the gene and its flanking regions.

Figure 10 shows the sequence of H1.10a compared to the originally sequenced H1.1a. Like most histone genes, neither gene contains an intron. H1.10a codes for a protein of 223 amino acids, six more than H1.1a. When the sequences of the two genes are aligned to accommodate the extra amino acids of H1.10a, the remainder of the two coding sequences are 90% homologous.

The differences between the two H1 sequences are not evenly distributed. Only one amino acid sequence difference (asn/ser) is in the central hydrophobic globular region of the molecule. This is also the region that is most conserved between species (4,5). The other 19 amino acid differences occur throughout the basic termini of the proteins. Most of the differences are nonconservative; that is, they involve amino acid side chains with significantly different chemical properties. The nonconservative differences (and also insertions) in H1.10a often involve

Figure 9. **Fine structure restriction map of p10a5.0BR. Labelling sites used for sequence determination (arrows) are shown.**

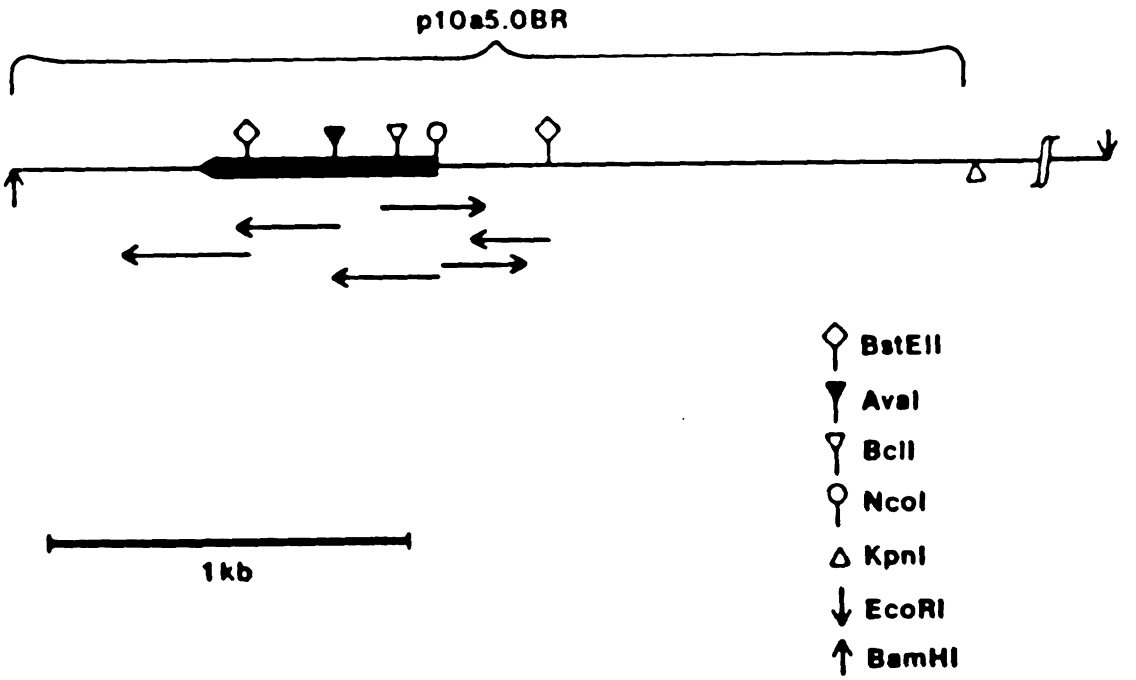


Figure 9

Figure 10. **Comparison of the sequences of H1.1a and H1.10a. The CCAAT element, TATAA element, coding sequence and 3' hyphenated dyad are capitalized. The H1 specific element is underlined. The proteins are different lengths; dashes are inserted in the sequence of H1.1a to align the genes to show homology. Amino acid differences are indicated.**

```
Cl11a     tggcggaatt gtagaaaaaa cgcgcttttt cgcctgttaa gaaacacaaa
Cl10a     ccaacgtccc ctcactcccg gtaaggaact ggctgcgcgg gcggtcaatt

          atagggggga gaagggagct ctgcgccgtg cgcgcggggc gggctctgca
          tctgtaggaa aagggatttt ttggaccgac aagaaacaca accggagcgg

          gcgcaCCAAT cacccgcggg ctccgctcTA TAAatacgag gccgccgact
          cggcggccgc gggccgaaCC AATcagcacg cgcgggcctg cTATAAaggg

          tgctccgggc ccagtggttc ccccgatct gtggaacgac gtccgtcacc
          ggcggcggcg gcgaccgcgc cattgttcgc ggagctccgc aggaggcgcc

              ser                                     ala
          ATG TCG GAG ACC GCG CCC GTT GCC GCG CCC GCG GTG TCT GCG
          ATG GCT GAG ACC GCT CCT GTC GCT GCG CCC GAT GTC GCC GCC
              ala                                     asp

          pro gly
          CCC GGC --- --- GCC AAG GCC GCC --- GCC AAG AAG CCG AAG
          GCC CCG ACC CCG GCC AAG GCG GCC CCC GCC AAG AAG CCG AAG
          ala pro thr pro                 pro

                                        pro
          AAG GCG GCG GGC GGC GCC AAG CCC CGC AAG CCC GCG GGC CCC
          AAG GCG GCG GGC GGC GCC AAA GCC CGC AAG CCC GCG GGC CCC
                                        ala


          AGC GTC ACC GAG CTG ATC ACC AAG GCC GTG TCC GCC TCC AAG
          AGC GTC ACC GAG CTG ATC ACC AAG GCC GTG TCC GCC TCC AAG



          GAG CGC AAG GGG CTC TCC CTC GCC GCG CTC AAG AAG GCG CTT
          GAG CGC AAG GGG CTC TCC CTC GCC GCG CTC AAG AAG GCG CTG


                                                asn
          GCC GCC GGC GGC TAC GAC GTG GAG AAG AAC AAC AGC CGC ATC
          GCC GCC GGC GGC TAC GAC GTG GAA AAA AGT AAC AGC CGC ATC
                                                ser


          AAG CTG GGG CTC AAG AGC CTC GTC AGC AAG GGC ACC CTG GTG
          AAG CTG GGG CTC AAG AGC CTC GTC AGC AAG GGC ACC CTG GTG
```

```
                    pro          thr                    thr
CH1a    AAA AAG CCG GGT GAG ACA AAA GAG AAA GCG ACT AAG AAG AAG
CH10a   AAG AAG TCG GGT GAT GTG AAG GAG AAG GCT CCT AAG AAG AAA
                    ser          val                    pro

        --- CCC GCG GCC AAG CCC AAG AAG CCG GCG GCC AAG AAG CCT
        ACG CCT GCA GCC AAG CCC AAG AAG CCG GCG GCC AAG AAA CCT
        thr

                                                   asn
        GCG GCT GCT GCC AAG AAG CCC AAG AAG GCA GCG GCG GTG AAG
        GCG GCT GCT GCC AAG AAG CCC AAG AAG GCG GTG GCA GTG AAG
                                                   val

        AAG AGC CCC AAG AAA GCC AAG AAG CCG GCA GCT GCT GCC ACC
        AAG AGC CCA AAG AAA GCT AAG AAG CCG GCG GCT GCC GCC ACC

                                                ala        gly
        AAG AAG GCG GCC AAG AGC CCC AAG AAG GCT ACC AAG GCT GGC
        AAG AAG GCG GCC AAG AGC CCC AAG AAG GTG ACC AAG GCT GCC
                                                val        ala

        arg          thr
        CGC CCC AAG AAG ACT --- GCC --- AAG AGC CCG GCC AAG GCA
        AAG CCC AAA AAG GCG GTG GCT GTC AAG AGC CCG GCC AAG GCA
        lys          ala val     val

                                                ser        ala
        AAG GCG GTG AAG CCC AAA GCT GCC AAG TCA AAG GCG GCC AAA
        AAG GCG GTG AAG CCC AAG GCT GCC AAG CCC AAG GCG ACC AAA
                                                pro        thr

                                                   thr
        CCC AAG GCG GCC AAG GCA AAG AAG GCA GCG ACC AAA AAG AAG
        CCC AAG GCA GCC AAG GCA AAG AAG GCG GCC CCT AAG AAA AAG
                                                   pro

        TAAgatgaca gaagaaattc gagtctgctc atttaaaaac cccaaaGGCT
        TAAatatcct ggggaaaaaa aaaaaaaaac cctcccctct gctttgcaga

        CTTTTAAGAG CCacccattt attctcagaa agagctggaa tgctgcggga
        aaaacccaac GGCTCTTTTA AGAGCCaccc aaagaaaccc aaaaagagcc

        accgcggcag cacaactaat tatctcagtt gcagagattc agatttgggc
        tttcgctgtt tcttacgatt ctttgtgtgt gtggagatgg aggttcgctt
```

while the basic character of the termini of H1.1a and H1.10a is the same, the different secondary structures of the termini may alter their functions (6).

The 5' nontranscribed sequences of genes are highly variable but contain short conserved sequences which act as control elements. The chicken H1 genes are no exception. Two elements which facilitate transcription are the TATAA and CCAAT elements (7,8). Both of these elements are found in many eukaryotic genes. The H1.10a gene, like the H1.1a gene, has both of these elements. The H1.10a TATAA element is 13 bp closer to the translation initiation codon ATG, so the untranslated leader of the H1.10a mRNA is somewhat shorter than that of the H1.10a mRNA. The location of the CCAAT element, relative to the TATAA element, is identical in the two genes. In each of these H1 genes the CCAAT sequence starts 23 bp upstream of the TATAA element.

A sequence element which has been found upstream of the CCAAT element in vertebrate H1 genes is an A+C-rich sequence containing the core sequence AAACACA (9). This sequence is found in both of these chicken H1 genes. A comparison of the A+C-rich elements from these genes shows an extended conserved sequence of AAGAAACACAA. However, the relative locations of the A+C-rich elements are different in these two genes. The element is 39 bp upstream of the CCAAT

element in H1.10A but 67 bp upstream of the H1.1a CCAAT element. The function of the conserved A+C-rich elements is not known.

The 3' ends of most eukaryotic mRNAs undergo post-transcriptional processing which adds a tail of polyadenosine. Most histone mRNAs are unusual in that they are not polyadenylated. Another type of processing generates specific 3' termini of histone mRNAs (10,11). This processing requires the presence of at least 2 histone gene-specific sequences: a hyphenated dyad which can form a stem+loop structure, and a purine-rich region 5-15 bp downstream of the dyad. These elements are thought to bind a small ribonucleoprotein during histone mRNA processing (12). The H1.1a histone gene was shown to have a hyphenated dyad sequence, GGCTCTTTTATAAGAGCC, which is similar or identical to the processing signal found in other histone mRNAs (1). The H1.10a gene contains the identical element which is flanked by sequences containing only A and C. This is characteristic of other histone genes as well. Fifteen bp downstream of each dyad is the start of a purine-rich sequence. The chicken element A(G/A)AAAGAG is similar to other vertebrate elements (13-15).

Other H1 gene sequences

All of the chicken H1 genes contain a BclI site

originating at these BclI sites were subcloned to facilitate sequence analysis of the other chicken H1 genes. This procedure also provided a rapid analysis of different H1 histone gene promoter regions, so that gene-specific S1 assays for mRNA levels could be designed.

Figure 11 shows the coding sequence of the 5' portions of the five chicken H1 genes. These portions of the genes code for the N-terminal basic domains of the histones. Each of the genes codes for a unique protein. Some differences between the H1 amino acid sequences are common to two or three variants.

H1.1a has a serine residue at the N-terminus. H1.1c differs from H1.1a at only one amino acid in this region of the protein. H1.2e differs from H1.1a at 5 positions, and has an inserted residue at position 20, but is like H1.1a and H1.1c by having serine at the N-terminus. The other two H1 genes, H1.10a and H1.2d, have alanine at the N-terminus. They also have 2 amino acids inserted after position 16 relative to H1.1a.

Figure 12 shows the promoter regions of the 5 chicken H1 genes. All of these promoters contain the TATAA element and the CCAAT element, though the H1.2e gene has a variant TATAA element, TAAAA. A comparison of the sequences flanking the prototypical TATAA elements shows that the

Figure 11. **Comparison of 5' coding sequences of chicken H1 genes. Dashes are inserted in some of the sequences to align the H1 protein sequences. Codons which specify nonconsensus amino acids are underlined. The overlapping SacII and ApaI sites found in all of the genes are indicated.**

```
H1.1a    ...ATG TCG GAG ACC GCG CCC GTT GCC GCG CCC GCG GTG TCT GCG CCC
H1.10a   ...ATG GCT GAG ACC GCT CCT GTC GCT GCG CCC GAT GTC GCC GCC GCC
H1.1c    ...ATG TCG GAG ACC GCT CCC GCC GCC GCC CCC GCT GTG TCC GCG CCC
H1.2d    ...ATG GCT GAG ACC GCT CCT GCT GCA GCG CCG GCG GCC --- GCG CCG
H1.2e    ...ATG GCC GAG ACC GCC CCC GCC GCC GCC CCC GAT GCG CCC GCG CCC

H1.1a       GCC --- --- GCC AAG GCC GCC --- GCC AAG AAG CCG AAG AAG GCG
H1.10a      GCG ACC CTG GCC AAG GCG GCC CCC GCC AAG AAG CCG AAG AAG GCG
H1.1c       GCC --- --- GCC AAG GCC GCC --- GCC AAG AAG CCG AAG AAG GCG
H1.2d       GCC GCC GCC GCC AAG GCC GCC --- GCC AAG AAG CCG AAG AAG GCG
H1.2e       GCC --- --- GCC AAG GCC GCC GCC GCC AAG AAG CCG AAG AAG GCG

H1.1a       GCG GCC GCC GCC AAG CCG CGC AAG CCC GCG GGC CCC AGC...BclI...
H1.10a      GCG GCC GCC GCC AAG GCC CGC AAG CCC GCG GGC CCC AGC...BclI...
H1.1c       GCG GCC GCC GCC AAG GCC CGC AAG CCC GCG GGC CCC AGC...BclI...
H1.2d       GCG GCC GCC GCC AAG GCC CGC AAG CCC GCG GGC CCG AGC...BclI...
H1.2e       GCG GCC GCC GCC AAG CCC CGC AAG CCC GCG GGC CCC AGC...BclI...

                                                CC GCG G = SacII
                                                   G GCC CC = ApaI
```

Figure 11

chicken H1 elements share a larger homology. The consensus sequence seen in these genes is C(T/G)CTATAAAT. This motif is similar to the element that is found in other histone genes (16).

The TATAA element directs the initiation of transcription to a site about 30 bp downstream of it. While gene-specific S1 analysis has not been performed to accurately determine the respective initiation sites of the H1 mRNAs, an estimate can be made using distance from TATAA and the fact that the start site is usually a purine in or 3' to a pyrimidine-rich region. These are shown in Figure 12. The location of the TATAA elements in the chicken H1 genes varies from 52 bp upstream of the initiation ATG codon (in H1.2d) to 82 bp upstream (in H1.1a). This means that all of the H1 mRNAs have rather short untranslated leaders.

The positioning of the CCAAT elements is quite uniform among the chicken H1 genes. The CCAAT elements are located 22, 23, or 24 bp upstream of the TATAA elements. This is closer than the positioning seen in other histone genes (16) and other eukaryotic genes generally (17). Usually CCAAT elements are found 40 to 60 bp upstream of TATAA elements. Like the H1 TATAA elements, the upstream H1 CCAAT elements have an extended homology. The larger sequence GCACCAATCA is found in 3 of the genes, and the elements

Figure 12. **H1 promoter sequences. CCAAT elements and TATAA elements are capitalized. All the sequences end at the initiation ATG. Probable mRNA cap sites are underlined.**

re 12

H1.1c and H1.2e contain the sequence CCAAT 25 bp downstream of the TATAA element. This is an anomolous position for a CCAAT element. Since these sequences are not within the larger consensus element and are not upstream of TATAA element, they probably do not have the regulatory activity of true CCAAT promoter elements.

At the time that this work was being done, Coles, Robbins, Madley, and Wells (19) reported the complete sequences of 4 chicken H1 histone genes, including the gene we were unable to isolate, to complete the characterization of the six genes in this family. Their results show that each gene codes for a unique protein, and that the conserved 5' and 3' elements we found in H1.10a are present in all the chicken H1 histone genes.

Since this group had completed the studies we had proposed, we did not continue our studies of the H1 histone family. We increased our efforts in an investigation of two other chromosomal protein genes, HMG-14 and HMG-17. The results of these studies are reported in Chapter 5 and the Appendix.

References

1. Sugarman, B., J.B. Dodgson, and J. Engel. (1983). J. Biol. Chem. 258:9005.

2. Grandy, D.G. (1986). Thesis. Michigan State University.

3. Ruiz-Carillo, A., M. Affolter, and J. Renaud. (1973). J. Mol. Biol. 170:843.

4. Allan, J., P.G. Hartman, C. Crane-Robinson, and F.X. Aviles. (1980). Nature (London) 288:675.

5. Carey, P.D., J.L. Hines, L.M. Bradbury, B.J. Smith, and E.W. Johns. (1981). Eur. J. Biochem. 120:371.

6. Bohm, L., and T.C. Mitchell. (1985). FEBS Lett. 193:1.

7. Parker, C.S., and J. Topol. (1984). Cell 36:357.

8. Jones, K.A., K.R. Yamamoto, and R. Tijian. (1985). Cell 42:559.

9. Coles, L.S., and J.R.E. Wells. (1985). Nuc. Acids Res. 13:585.

10. Busslinger, M., D. Schumperli, and M.L. Birnstiel. (1985). Cold Spring Harbor Symp. Quant. Biol. 50:665.

11. Birnstiel, J.L., M. Busslinger, and K. Strub. (1985). Cell 41:349.

12. Strub, K., G. Galli, M. Busslinger, and M.L. Birnstiel. (1984). EMBO J. 3:2801.

14. Perry, M., G.H. Thomsen, and R.G. Roeder. (1985). J. Mol. Biol. 185:479.

15. Mezquita, J., W. Conner, R.J. Winkfein, and G.H. Dixon. (1985). J. Mol. Evol. 21:209.

16. Hentschel, C.C., and M.L. Birnstiel. (1981). Cell 25:301.

17. McKnight, S.L., and R. Kingsbury. (1982). Science 217:316.

18. D'Andrea, R.J., L.S. Coles, C. Lesnikowski, L. Tabe, and J.R.E. Wells. (1985). Mol. and Cell. Biol. 5:3108.

19. Coles, L.S., A.J. Robbins, L.K. Madley, and J.R.E. Wells. (1987). J. Biol. Chem. 262:9656.

Chapter 4

Literature Review: HMG Chromosomal Proteins

The structure of chromatin and the regulation of expression of its information is determined by complicated interactions between nucleic acids and proteins. The most abundant of the chromosomal proteins are the histones (see Chapter 1). Nonhistone chromosomal proteins range from proteins of low abundance and high specificity of action, perhaps only transiently associated with chromatin, to proteins which are nearly as abundant as histones. The most abundant of the nonhistone chromosomal proteins are the high mobility group proteins (HMG proteins). There are approximately 10^6 molecules of each type of HMG protein in most cells, so they are about one-tenth as abundant as the histones (1).

HMG proteins were first characterized by E.W. Johns and his colleagues. They were studying the histones of calf thymus chromatin. To isolate histone H1 they started by extracting tissue with perchloric acid (2). This extract contained H1 as its only histone component, but also contained upwards of 20 other proteins at measurable levels. Most of these proteins can be released from isolated chromatin by 0.35 M NaCl washes (3) which are too mild to extract histones. Because these proteins migrated rapidly in the polyacrylamide acid-urea gels used to separate basic chromosomal proteins like histones, they were named HMG-1,

products of histones or larger HMG proteins. HMG-3, for example, is derived from the N-terminal portion of HMG-1 (4). Levels of HMG-1 decrease when HMG-3 increases in a preparation. Another persistent artifact of preparation is HMG-8, which is derived from the N-terminal portion of histone H1 (5). Nuclear proteases are probably responsible for the occurrences of many of these HMG fractions. When efforts to diminish proteolysis during preparation are successful, four major HMG proteins are found. It is now well established that HMG-1, HMG-2, HMG-14, and HMG-17 are the major HMG proteins in vertebrate tissues. There appear to be 2 subgroups: HMG-1 and HMG-2 are similar, and HMG-14 and HMG-17 are similar.

The 4 HMG proteins of calf thymus have characteristics which have been used to define this type of protein.

1) They are chromosomal proteins.
2) They can be extracted from chromatin by 0.35M NaCl.
3) They are small proteins, less than 30 kd.
4) They are rich in lysine and arginine (>20%).
5) They are rich in glutamic acid and aspartic acid (>20%).
6) They are rich in proline (>5%).

Proteins very like the calf thymus HMG proteins have

(9-12). Primary sequence information confirms the homology of some of these proteins to the prototypical calf thymus proteins.

The occurrence of HMG proteins in nonvertebrate eukaryotes is not well documented. An early report described proteins isolated from yeast (13) and wheat (14) chromatin which fit the physical criteria of Johns et al., but without sequence information, no assignment of homology could be made. These proteins are intermediate in size between calf HMG-1,2 and HMG-14,17. Proteins with HMG-like size and amino acid composition have been isolated from several types of insect chromatin, including Drosophila melanogaster (15,16), D.virilis (17), and Ceratitus capitata (18). Again, the sequences of these proteins are not known so they are only nominally HMG proteins. The strongest evidence for an HMG protein outside of those of vertebrates is a recent report of an HMG-1,2 homologue in Saccharomyces cervisiae called ACP2. The sequence of the ACP2 gene was determined and the deduced amino acid sequence matches the calf HMG1 sequence at 43% of their residues when conservative replacements are included (19). This level of homology indicates that ACP2 is indeed a member of the HMG-1,2 family, since the trout and calf proteins have only 58% identity. Disruption of ACP2 gene function is lethal. The

fact, homologues of the vertebrate proteins. If, as many assume, the HMG proteins play a fundamental role in the dynamics of chromatin structure, this would be expected. Absent a functional assay for any HMG protein, homologous protein sequences from other phylogenetically diverse species would be reassuring.

HMG-1 and HMG-2

HMG-1 and HMG-2 are similar in size (26 kd) and amino acid composition (20). Comparison of the amino acid sequences that are known confirms that these two proteins are closely related and they are sometimes collectively known as HMG-1,2. The primary sequences of a number of vertebrate HMG-1,2 proteins have been determined. Figure 13 shows an example from a rat. The composition of this HMG-1 protein is typical of the family: lys+arg = 24%, asp+glu = 27%, and pro = 6%. The charges are not evenly distributed throughout the molecule. The C-terminus consist of a run of 30 acidic residues, while the N-terminal two-thirds of the molecule is basic.

On the basis of their amino acid sequences (21) a 3-domain structure for the HMG-1,2 proteins has been proposed (22). Domains A and B are globular basic regions with moderate helicity (23). These are residues 1 to 79 and residues 90 to 163 of rat (and human). A 10 residue linker

Figure 13. **The sequence of a rat HMG-1 cDNA. The C-terminus of the predicted protein is made only of lysine (K), aspartic acid (D), and glutamic acid (E). (From (24)).**

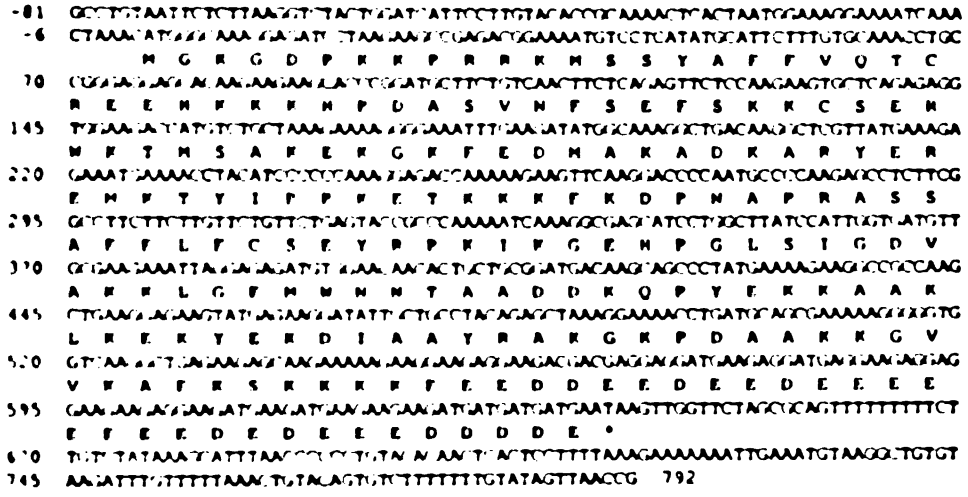


Figure 13

segment that precedes it. The high charge of this domain prevents the formation of stable structures in solution.

Microheterogeneity

When preparations of HMG proteins are analyzed by electrophoresis or high pressure liquid chromatography, microheterogeneity is found in every HMG protein fraction (25-28). The reason for this is that all of the HMG proteins undergo a variety of posttranslational modifications, much like the histones. The functions of these modifications are not known for either histones or HMG proteins.

When purified calf thymus HMG-1 and HMG-2 proteins are analyzed, 3-4% of HMG-1 and 8-9% of HMG-2 is found to be methylated (29). The modified residue is N^6-N^6-dimethyl arginine, but which specific arginines are modified is not known. No methylated lysine residues are found. HMG-1,2 can also be acetylated (30-32). The modified residues are lysines in the N-terminal 12 amino acids of the proteins which are released by treatment with cyanogen bromide. Sterner et al. (32) added ^{3}H-acetate to a calf thymus homogenate to radioactively label the modified residues. When they sequenced the N-terminal peptide fragment they found only 2 of the 4 lysines it contains were labelled, specifically those at positions 2 and 11. The N-terminus

Several types of glycosylation have been demonstrated (33). The sugars mannose, galactose, glucose, fucose, N-acetylglucosamine, and an unidentified sugar (possibly xylose) can be released from HMG-1,2 by alkaline borohydride hydrolysis. Mild alkaline hydrolysis, which releases O-linked sugars, will not release the carbohydrate from HMG proteins, suggesting they are N-glycosyl linkages. HMG-1,2 are reportedly modified by ADP-ribosylation, but neither the site nor the structure of this modification is well characterized (34).

Possible HMG-1,2 functions

Despite many investigations, no clear functions for the HMG proteins are known. Studies of HMG protein function have generally been of 2 types: 1) since the HMG proteins are major components of chromatin, their interactions with DNA, histones, and nucleosomes have been studied, and 2) the HMG proteins, or nucleosomes associated with them, have been characterized from various cell types.

One of the first properties of HMG-1,2 discovered was that the molecules have a much higher affinity for single stranded DNA than for double stranded DNA. This is demonstrated in the most straightforward way by passing HMG proteins over columns which contain immobilized DNA.

not (35). It has been suggested that one function of the HMG proteins might be destabilizing the DNA double helix by virtue of their affinity for single stranded DNA.

Yoshida and his colleagues described an interesting interaction of HMG-1,2 with supercoiled DNA. They inserted a $(CG)_{10}$ fragment into the ampicillin resistance gene of pBR322. When this plasmid is highly supercoiled the 20 bp insert assumes the Z-DNA form. When this plasmid is put into an *in vitro* transcription reaction, *E. coli* RNA polymerase cannot transcribe through the region of Z-DNA and an abbreviated transcript is made, but when HMG-1 is included in the reaction the transcriptional block is removed (36). They think that HMG-1 binds in or near the region of Z-DNA and relieves supercoiling there, even as it increases supercoiling in the rest of the plasmid.

Another interaction of HMG-1 and DNA was recently described by Bianchi *et al.* (37). They designed DNA fragments which anneal into cruciform DNA, made an affinity chromatography column with these cruciform fragments, and purified a cruciform binding protein from rat liver. Peptide analysis and sequencing showed that this protein is in fact HMG-1,2. *Bona fide* rat HMG-1, produced from a cloned cDNA, binds cruciform DNA as measured by a gel shift assay. Because of the clever design of their artificial

do not bind HMG-1, and suggest the HMG-1 binds to the branch point where one double strand splits to two single strands. This is consistent with all but the earliest results described above.

Nucleosomes can be reconstituted from their component DNA and histones, but only when the components are mixed in a high salt solution, followed by lengthy dialysis to physiological conditions (38,39). The reassembly of nucleosomes at physiological ionic strengths can be facilitated by the addition of certain Xenopus oocyte extracts (40). When the "assembly factor" is further purified, it is found to be an acidic protein (41) called nucleoplasmin. Nucleoplasmin is the predominant nucleoprotein of Xenopus oocytes and in vitro it appears to act by preventing nonspecific aggregate formation between DNA and the large pools of histones present in the oocytes, thus allowing proper nucleosome assembly. Acidic polypeptides (polyglutamic acid or polyaspartic acid) can substitute for nucleoplasmin in the assembly reaction (42). These facts led Bonne-Andre et al. to investigate the possibility that HMG-1, with its acidic C domain and its histone-interacting A domain, might be a nucleosome assembly factor (43). They found that HMG-1 facilitiates the formation of core octamers of histones. Also, the addition

the reaction. The nucleosomes reconstituted in the presence of HMG-1 appear to be normal in electron micrographs.

It is not yet possible to decide which of the properties of HMG-1,2 are biologically relevant. Whether HMG-1,2 is involved in DNA binding or unwinding or melting or nucleosome assembly, a role in either of the two basic functions of chromatin, replication and transcription, can be postulated.

HMG-14 and HMG-17

Originally, two small (9-12 kd) HMG proteins were found in mammalian and avian cells. These are HMG-14 and HMG-17 (44). The amino acid compositions of these proteins are similar and show that these proteins form a small family, so they are sometimes collectively referred to as HMG-14,17. Trout have a single small HMG protein, called H6 (45). It is a member of the HMG-14,17 family, but because of its amino acid composition and small size (7 kd) its relation to individual mammalian HMG proteins is not clear.

Several of the small HMG proteins have been completely sequenced, including calf HMG-17 (46), calf HMG-14 (47), chicken HMG-17 (48), and trout H6 (49), and a partial sequence of chicken HMG-14 is known (50). The sequences clearly show similarities throughout the proteins, confirming that these proteins are related (51).

HMG-14,17 contain only a low proportion of hydrophobic amino acids. Structural studies of HMG-14 (52) and HMG-17 (53) using circular dichroism and nuclear magnetic resonance spectroscopy show that these proteins have little or no secondary or tertiary structure in a wide range of solution conditions. These methods only detect significant structural involvement of amino acid side chains when the proteins interact with DNA at low ionic strength. The residues involved in this DNA binding are in the N-terminal parts of the molecules, approximately between residues 15 and 40. This region is highly conserved in all of the HMG-14,17 molecules examined.

HMG-14,17, like HMG-1,2, undergo a number of post-translational modifications which cause microheterogeneity when protein preparations are analyzed by electrophoresis or, especially, high pressure liquid chromatography. The best studied of these is phosphorylation. In vitro studies show that mammalian HMG-14,17 can be substrates for phophorylation by cAMP- and cGMP-dependent kinases (54,55) and other kinases (56,57). Both serine and threonine side chains can be phosphorylated in vitro. Several researchers have tried to correlate in vivo levels of HMG-14,17 phosphorylation to stages of the cell cycle by adding ^{32}P-

studies show that phosphorylation of HMG-14 is higher in metaphase-arrested HeLa cells than cells in interphase (58-60). An endogenous kinase of chromatin may be responsible for this phosphorylation. Another study shows that an HMG-14-like protein is phosphorylated in both metaphase and interphase cells, but that the electrophoretic mobility of the metaphase protein is less (61). Multiply phosphorylated forms of HMG-14 may account for the higher levels of HMG-14 phosphorylation seen in all the studies; basal levels of HMG phosphorylation may not be detected by some labelling regimens.

Phosphorylation of HMG-17 was not detected in the work just described, but has been reported in other studies. Both HMG-14 and HMG-17 are phosphorylated in Chinese hamster ovary cells in interphase. When these cells are arrested at metaphase, the relative phosphorylation of HMG-14 increases (62,63). Phosphorylated HMG-17 has also been found in rat cells (64) and mouse cells (65).

HMG-14,17 can be acetylated at lysine residues in the N-terminal portion of the molecules (65,66). Acetylation of chromosomal proteins may have profound effects on gene structure and activity (67). In vitro, HMG-14,17 can inhibit the action of histone deacetylase (68,69). This property may be important for their in vivo function since

Mammalian HMG-14,17 can also be modified by a number of carbohydrate moieties. Fructose, galactose, mannose, and N-acetyl glucosamine have been found in these proteins, but it is not known what percentage of them is modified or how heterogeneous the glycosylated forms are (71). ADP-ribosylation of HMG-14,17 has been reported in transformed human cells (72) and mouse mammary carcinoma cells (73). Only 0.03% of the HMG proteins are ADP-ribosylated in the mouse cells, but this small fraction may play a regulatory role since glucocorticoid treatment, which induces transcription of some cellular and tumor virus genes, decreases the level of HMG ADP-ribosylation significantly (74).

Possible HMG-14,17 functions

During development of an organism, specific sets of genes are transcribed in various cells at various times. Genes that are to be activated are assembled into a specific chromatin structure prior to the onset of their transcription (75). Despite considerable efforts, the unique components and structure of transcriptionally active chromatin are not well understood.

The best evidence that active chromatin is structurally distinct from inactive chromatin has come from nuclease digestion experiments. These experiments show that active

sensitivity is seen throughout the region of an active gene (77,78). Superimposed upon this generally increased nuclease sensitivity, many active genes have sites which are hypersensitive to DNase I or micrococcal nuclease (77). These hypersensitive sites are often in the control sequences 5' to active genes. The appearance of hypersensitive sites before transcription begins and their persistence after transcription ceases show that these sites are indications of an activated chromatin structure, rather than a consequence of the process of transcription itself (78,79). The active chromatin structure indicated by DNase I hypersensitivity is necessary but not sufficient for transcription induction in most genes that have been examined.

Light digestion of chromatin by nucleases releases mononucleosomes and oligonucleosomes which are substantially enriched in actively transcribed sequences (75,80). Studying the components of these nucleosomes, and reconstituting nuclease-sensitive chromatin from isolated components, has provided evidence that HMG-14 and HMG-17 are involved in the maintenance of a transcriptionally active chromatin structure.

The earliest report relating HMG-14 and HMG-17 to nuclease sensitivity was by Weisbrod and Weintraub (81).

erythrocyte chromatin but not in brain chromatin. This sensitivity is lost when erythrocyte chromatin is extracted with 0.35 M NaCl, which removes some chromosomal proteins including HMG-14 and HMG-17. Reconstitution of the depleted erythrocyte chromatin with HMG-14 or HMG-17 restores DNase I sensitivity, even when brain is the source of the HMG proteins. This shows that, though HMG-14 and/or HMG-17 are a necessary part of the active globin nucleosomes, some other feature of the nucleosome is responsible for tissue-specific transcription patterns.

Weisbrod *et al*. extended this work to show that most of the active genes of erythrocytes and of a leukemia cell line have nuclease sensitivity which is conferred by HMG-14 and HMG-17 (82). By measuring the hybridization kinetics of nuclease-generated nucleosomes and total nuclear RNA they could demonstrate the involvement of HMG-14 and HMG-17 throughout the active portion of the genome. The bulk stoichiometry of their reconstitution experiments is puzzling. Although 20% of the genome is transcribed in these cells, nuclease sensitivity is restored to individual genes (*e.g.*, β- globin) by just one mole of either HMG-14 or HMG-17 protein per 20 moles of nucleosomes. The selective affinity of HMG-14 and HMG-17 for globin nucleosomes in erythrocytes was confirmed by Sandeen *et al*. (83). They

nucleosomes have 2 strong binding sites for either HMG-14 or HMG-17, confirming other studies (84,85). They also showed that HMG-14 and HMG-17 bind naked DNA, albeit less tightly.

In the studies cited above, active nucleosomes were shown to be associated with HMG-14,17. Two groups have taken the reverse approach by isolating HMG-associated nucleosomes and characterizing the sequences they contain. The results of these studies are mixed. Dorbic and Witting (86) used a monoclonal anti-HMG-17 antibody to isolate nucleosomes produced by light nuclease digestion of nuclei. They found that nucleosomes released from liver nuclei contain a gene active in liver (vitellogen II) and oviduct nucleosomes contain genes active in oviduct (ovalbumin and lysozyme), but not <u>vice versa</u>. Druckmann <u>et al</u>. also used an antibody to enrich a fraction of HMG-17-containing nucleosomes from the livers of rats before or after treatment with a carcinogen (87). This carcinogen induces a P450 liver enzyme. They examined the HMG-17 enriched fraction for the presence of repetitive DNA, non-transcribed DNA, transcribed genes, and the inducible P450 gene. They concluded that actively transcribed genes are enriched in the fraction of nucleosomes which contain HMG-17; but, to a lesser extent, so are some genes that are not actively transcribed.

The nucleosomes used in the study described first were prepared by light digestion of nuclei, whereas the nucleosomes in the second study were prepared from a much more complete digest of prepared chromatin. It seems likely that higher levels of digestion of chromatin may destroy some higher orders of chromatin structure. Other studies have implicated higher order structures in the maintenance of nuclease sensitivity (88). It is possible that some rearrangement of chromatin components occurs during chromatin isolation so that the structure of active chromatin is obscured when chromatin is isolated. Furthermore, if HMG-14,17 need not associate with every nucleosome in an active array, complete digestion to mononucleosomes before antibody binding will lessen the specific enrichment for active genes. These problems are not limited to studies of HMG-14,17 function; they constitute a virtually unavoidable problem in relating biochemical observations of solubilized chromatin to the roles of histones, HMG proteins and other chromosomal proteins that constitute the active chromosomal DNA structures in the living cell.

To allow further study of the function, production, and evolution of HMG-14 and HMG-17 we have isolated HMG-14 and HMG-17 cDNA clones and genomic clones. At the start of this

isolation of human HMG-14 cDNA clones (89) and HMG-17 cDNA clones (90) in 1986. These clones were constructed in _gt11, a vector which allows expression of the cloned cDNA, so that anti-HMG antibodies could be used to isolate the HMG cDNA clones. When the HMG cDNA clones were used as probes to study the HMG genes it was found that both the human HMG-14 and HMG-17 genes are members of multigene families with about 50 members; most members of these families are processed retropseudogenes (91). They isolated 125-150 genomic clones by hybridization to the cDNA clones. The active genes were isolated from this set by hybridization to a series of oligonucleotides complementary to the 3' portions of the cDNAs (92,93). They also used the human HMG-14 and HMG-17 cDNA clones as probes to isolate a chicken HMG-14 cDNA clone (94), a chicken HMG-17 cDNA clone (95) and the chicken HMG-17 gene (96). The results their studies of these clones are compared to ours in Chapter 5.

References

1. Gabrielli, F., R. Hancock, and A.J. Faber. (1981). Eur. J. Biochem. 120:363.
2. Sanders, C., E.W. Johns. (1974). Biochem. Soc. Trans. 2:547.
3. Goodwin, G.H., J.M. Walker, E.W. Johns. (1978). Biochim. Biophys. Acta 519:233.
4. Johns, E.W. (1967). Biochem. J. 104:78.
5. Chapman, G.E., P.G. Hartman, and E.M. Bradbury. (1976). Eur. J. Biochem. 61:69.
6. Faire, R.J., and D.W. Cooper. (1987). Comp. Biochem. Physiol. B Comp. Biochem. 87:423.
7. Walker, J.M., and E.W. Johns. (1980). Biochem. J. 185:383.
8. Walker, J.M., C. Stearn, and E.W. Johns. (1980). FEBS Lett. 112:207.
9. Watson, D.C., E.H. Peters, and G.H. Dixon. (1977). Eur. Journ. Biochem. 74:53.
10. Watson, D.C., N.C.W. Wong, and G.H. Dixon. (1979). Eur. Jour. Biochem. 95:193.
11. Rabbani, A., G.H. Goodwin, J.M. Walker, E. Brown, and E.W. Johns. (1980). FEBS Lett. 109:294.
12. Kennedy, B.P., and P.L. Davies. (1980). J. Biol. Chem. 255:2533.

14. Spiker, S., J.K.W. and I. Isenberg. (1978). Biochem. Biophys. Res. Com. 82:129.

15. Alfageme, C.R., G.T. Rudkin, and L.H. Cohen. (1980). Chromosoma 78:1.

16. Bassuk, J.A., and J.E. Mayfield. (1982). Biochemistry 21:1024.

17. Krasnov, P.A., A.A. Karavanov, and L.I. Lorochkin. (1984). Ontogenez 15:547.

18. Marquez, G., F. Moran, L. Franco, and F. Montero. (1982). Eur. J. Biochem. 123:165.

19. Haggren, W., and D. Kolodrubetz. (1988). Mol. Cell Biol. 8:1282.

20. Walker, J.M. (1982). in "HMG Chromosomal Proteins", Johns, E.W., ed., p. 223. Academic Press, New York.

24. Paonessa, G., R. Frank, and R. Cortese. (1987). Nuc. Acids Res. 15:9077.

21. Chou, P.Y., and G.D. Fasman. (1978). Adv. in Enzymology 47:45.

22. Reeck, G.R., P.J. Isackson, and D.C. Teller. (1982). Nature 300:76.

23. Land, D., D.J. Cox, D.R. Manning, and G.R. Reeck. (1985). Biochim. Biphys. Acta 831:207.

25. Elton, T.S., and R. Reeves. (1985). Anal. Biochem. 144:403.

27. Elton, T.S., and R. Reeves. (1986). Anal. Biochem. 157:53.

28. Goodwin, G.H., P.N. Cockerill, S.Kellam, and C.A. Wright. (1985). Eur. J. Biochem. 149:47.

29. Boffa, L.C., R. Sterner, G. Vidali, and V.G. Allfrey. (1979). Biochem. Biophys. Res. Com. 89:1322.

30. Nicolas, R.N., and G.H. Goodwin. (1982). in "The HMG Chromosomal Proteins", Johns, E.W., ed., p. 41. Academic Press, New York.

31. Sterner, R., G. Vidali, R.L. Heinrickson, and V.G. Allfrey. (1978). J. Biol. Chem. 253:7601.

32. Sterner, R., G. Vidali, and V.G. Allfrey. (1979). J. Biol. Chem. 254:11577.

33. Reeves, R., D. Chang, and S.C. Chung. (1981). Proc. Nat. Acad. Sci. USA 78:6704.

34. Poirier, G.G., C. Niedergang, M. Champagne, A. Mazen, and P. Mandel. Eur. J. Biochem. 127:437.

35. Isackson, P.J., J.L. Fishback, D.L. Bidney, and G.R. Reeck. (1979). J. Biol. Chem. 254:5569.

36. Waga, S., S. Mizuno, and M. Yoshida. (1988). Biochem. Biophys. Res. Com. 153:334.

37. Bianchi, M.E., M. Beltrame, and G. Paonessa. (1989). Science 243:1056.

38. Oudet, P., M. Gross-Bellard, and P. Chambon. (1975).

39. Camerini-Otero, R.D., B. Sollner-Webb, and G. Felsenfeld. (1976). Cell 8:333.

40. Laskey, R.A., A.D. Mills, and N. Morris. (1977). Cell 10:237.

41. Laskey, R.A., B.M. Honda, A.D. Mills, and J.T. Finch. (1978). Nature (London) 275:416.

42. Stein, A., J.P. Whitlock, and M. Bina. (1979). Proc. Nat. Acad. Sci. USA 76:5000.

43. Bonne-Andrea, C., F. Harper, J. Sobczak, and A.-M. De Recondo. (1984). EMBO J. 3:1193.

44. Goodwin, G.H., R.H. Nicolas, and E.W. Johns. (1975). Biochim. Biophys. Acta 405:280.

45. Dixon, G.H. (1978). in "The HMG Chromosomal Proteins", Johns, E.W., (ed.), p. 149. Academic Press, New York.

46. Walker, J.M., J.R.B. Hastings, and E.W. Johns. (1977). Eur. J. Biochem. 76:461.

47. Walker, J.M., G.H. Goodwin, and E.W. Johns. (1979). FEBS Lett. 100:394.

48. Walker, J.M. (1982). in "The HMG Chromosomal Proteins", Johns, E.W., (ed.), p. 69. Academic Press, New York.

49. Watson, D.C., N.C.W. Wong, and G.H. Dixon. (1979).

50. Walker, J.M., E. Brown, G.H. Goodwin, C. Stearn, and E.W. Johns. (1980). FEBS Lett. 113:253.

51. Reeck, G.R., and D.C. Teller. (1985). in "Progress in Nonhistone Protein Research", Vol. II, Bekhor, I., (ed.), p. 10. CRC Press, Boca Raton.

52. Javaherian, K., and S. Amini. (1978). Biochem. Biophys. Res. Com. 85:1385.

53. Abercrombie, B.D., G.G. Kneale, C. Crane-Robinson, E.M. Bradbury, G.H. Goodwin, J.M. Walker, and E.W. Johns. (1978). Eur. J. Biochem. 84:173.

54. Palvimo, P., A. Linnala-Kankkunen, and P.H. Maenpaa. (1983). Biochem. Biphys. Res. Com. 110:378.

55. Taylor, S.S. (1982). J. Biol. Chem. 257:6056.

56. Harrison, J.J., and R.A. Jungman. (1982). Biochem. Biophys. Res. Com. 108:1204.

57. Inoue, A., Y. Tei, T. Hasuma, M. Yukioka, and S. Morisawa. (1980). FEBS Lett. 117:68.

58. Walton, G.M., and G.N. Gill. (1983). J. Biol. Chem. 258:4440.

59. Bhorjee, J.S., I. Mellon, and L. Kifle. (1983). Biochem. Biophys. Res. Com. 111:1001.

60. Paulson, J.R., and S.S. Taylor. (1982). J. Biol. Chem. 257:6064.

61. Lund, T., J. Hotland, M. Fredricksen, and S.G. Laland.

62. Arfmann, H.A., and H. Baydoun. (1981). Z. Naturforsch. 36:319.

63. Arfmann, H.A., E. Haase, and H. Schroter. (1981). Biochem. Biophys. Res. Com. 101:137.

64. Harrison, J.J., and R.A. Jungmann. (1982). Biochem. Biophys. Res. Com. 108:1204.

65. Sterner, R., G. Vidali, and V.G. Allfrey. (1979). J. Biol. Chem. 254:11577.

66. Sterner, R., and V.G. Allfrey. (1983). J. Biol. Chem. 258:12135.

67. Allfrey, V.G. (1980). in "Cell Biology. A Comprehensive Treatise", Goldstein, L. and D. Prescott, (eds.), p. 347. Academic Press, New York.

68. Reeves, R., and E.P. Candido. (1980). Nuc. Acids Res. 8:1947.

69. Mezquita, J., J. Chiva, S. Vidal, and C. Mezquita. (1982). Nuc. Acids Res. 10:1781.

70. Malik, N., M. Smulson, and M. Bustin. (1984). J. Biol. Chem. 259:699.

71. Reeves, R., D. Chang, and S.C. Chung. (1981). Proc. Nat. Acad. Sci. USA 78:6704.

72. Giri, C.P., M.H. West, and M. Smulson. (1978). Biochemistry 17:3495.

73. [illegible] and C.S. Johnson. (1983). J. Biol. Chem.

74. Tanuma, S., L.D. Johnson, and G.S. Johnson. (1983). J. Biol. Chem. 258:15371.

75. Weintraub, H., and N. Groudine. (1976). Science 193:848.

76. Garel, A., and R. Axel. (1976). Proc. Nat. Acad. Sci. USA 73:3966.

77. Weisbrod, S. (1982). Nature (London) 297:289.

78. Wood, W.I., and G. Felsenfeld. (1982). J. Biol. Chem. 257:7730.

79. Groudine, M., and H. Weintraub. (1982). Cell 30:131.

80. Levy, W.B., and G.H. Dixon. (1978). Nuc. Acids Res. 5:4155.

81. Weisbrod, S., and H. Weintraub. (1979). Proc. Nat. Acad. Sci. USA 76:630.

82. Weisbrod, S., M. Groudine, and H. Weintraub. (1980). Cell 19:289.

83. Sandeen, G., W.I. Wood, and G. Felsenfeld. (1980). Nuc. Acids Res. 8:3757.

84. Albright, S.C., J.M. Wiseman, R.A. Lange, and W.T. Garrad. (1980). J. Biol. Chem. 255:3673.

85. Jackson, J.B., J.M. Pollock, and R.L. Rill. (1979). Biochemistry 18:3739.

86. Dorbig, T., and B. Wittig. (1986). Nuc. Acids Res.

87. Druckmann, S., E. Mendelson, D. Landsman, and M. Bustin. (1986). Exp. Cell Research 166:486.

88. Senear, A.W., and R.D. Palmiter. (1981). J. Biol. Chem. 256:1191.

89. Landsman, D., T. Thyagarajan, R. Westermann, and M. Bustin. (1986). J. Biol. Chem. 261:16082.

90. Landsman, D., N. Soares, F.J. Gonzalez, and M. Bustin. (1986). J. Biol. Chem. 261:7479.

91. Srikantha, T., D. Landsman, and M. Bustin. (1987). J. Mol. Biol. 197:405.

92. Landsman, D., O.W. McBride, N. Soares, M.P. Crippa, T. Srikantha, and M. Bustin. (1989). J. Biol. Chem. 264:3421.

93. Landsman, D., O.W. McBride, and M. Bustin. (1989). Nuc. Acids Res. 17:2301.

94. Srikantha, T., D. Landsman, and M. Bustin. (1988). J. Biol. Chem. 263:13500.

95. Landsman, D., and M. Bustin. (1987). Nuc. Acids Res. 15:6750.

96. Landsman, D., T. Srikantha, and M. Bustin. (1988). J. Biol. Chem. 263:3917.

Chapter 5

Isolation and Characterisation of Chicken HMG-14a and HMG-17 Genes

The Appendix describes the isolation of cDNA clones derived from chicken HMG-14 and HMG-17 mRNAs. The known amino acid sequences of HMG-14 and HMG-17 were used to predict possible mRNA sequences. A mixture of oligonucleotides which would hybridize to these mRNA sequences was used successfully to screen a λgt11 cDNA library generated from chicken liver mRNA (1). DNA sequence analysis showed that pLM3b is a HMG-14 cDNA, and pLG1a is a HMG-17 cDNA. When the HMG cDNAs were used as hybridization probes of chicken genomic Southern blots, the HMG-14 and HMG-17 genes were judged to be single copy genes.

Srikantha, Landsman, and Bustin (2) used a different approach to isolate a chicken HMG-14 cDNA. They used a human HMG-14 cDNA to isolate homologous chicken cDNAs and determined the sequence of one of them. The sequence of their HMG-14 cDNA is quite different from the one we isolated, and the cDNAs detect different single copy genes with Southern hybridization regimens of normal stringency. Though the cDNA they isolated is clearly a member of the HMG-14 family, the amino acid composition of the protein it encodes differs significantly from the amino acid composition of the major chicken HMG-14 protein. In particular, amino acid analysis shows that chicken HMG-14 has 10-11 prolines, 5 serines, 1 histidine, and no valine. While the cDNA we isolated encodes a protein with this

encodes a protein with 8 prolines, 8 serines, no histidine, and 5 valines. Since the discovery of this second member of the chicken HMG-14 family, the cDNA we isolated (and its gene) has been called HMG-14a, while the human homologue has been called HMG-14b.

It is not yet known if mammals contain a homologue of chicken HMG-14a. Whether or not a human HMG-14a homologue can be isolated by nucleic acid hybridization, the homology of HMG-14b and the human transcript suggests that the HMG-14 family is an ancient one. The HMG-14b gene has been isolated and characterized by Srikantha, Landsman, and Bustin (3). The structure of HMG-14b is compared with HMG-14a below.

The chicken HMG-17 gene

To isolate the chicken HMG-17 gene, the chicken HMG-17 cDNA pLG1a was used as a radioactive hybridization probe in a screen of a λCharon4A library of chicken genomic DNA clones as described in Chapter 2. Nine plaques tested positive in the initial hybridization screen. When these 9 clones were isolated, replated, and rescreened, 3 of them gave strong positive signals. These were λHM5a, λHM6b, and λHM7a. DNA was prepared from the purified clones and used for restriction mapping and Southern analysis.

Figure 14 shows restriction maps of λHM5a, λHM6b, and

blots were hybridized to the HMG-17 cDNA probe to identify fragments which contain coding sequences of the HMG-17 gene. The resultant hybridization patterns also facilitated alignment of the maps to indicate overlap among these clones. λHM7a appeared to contain the entire HMG-17 coding region.

Two fragments of λHM7a hybridize strongly to the HMG-17 cDNA probe (Figure 14). These fragments, a 1.7 kb BamHI-HinDIII fragment and a 2.1 kb HinDIII fragment, were isolated and subcloned into a plasmid vector (Bluescribe KS+). The resulting subclones, pHM1.7BH and pHM2.1HH, were used for restriction mapping and sequence analysis. (For reasons discussed below, it also became necessary to subclone the 0.8 kb HinDIII-BamHI fragment which adjoins pHM1.7BH. This fragment does not hybridize to the HMG-17 cDNA probe.)

One immediate conclusion that could be drawn from the initial mapping and Southern analysis of λHM7a was that the HMG-17 gene contains an intron. The HMG-17 cDNA contains no HinDIII sites, yet 2 HinDIII fragments hybridize to the HMG-17 cDNA probe. This suggested that the HinDIII site between pHM1.7BH and pHM2.1HH lies within an intron.

The previously characterized HMG-17 cDNA sequence enabled us to predict restriction sites which should be found in the exons of the HMG-17 gene. These enzyme sites

Figure 14. **Restriction maps of recombinant λ Charon4A chicken genomic clones which hybridize to the HMG-17 cDNA pLG1a.**

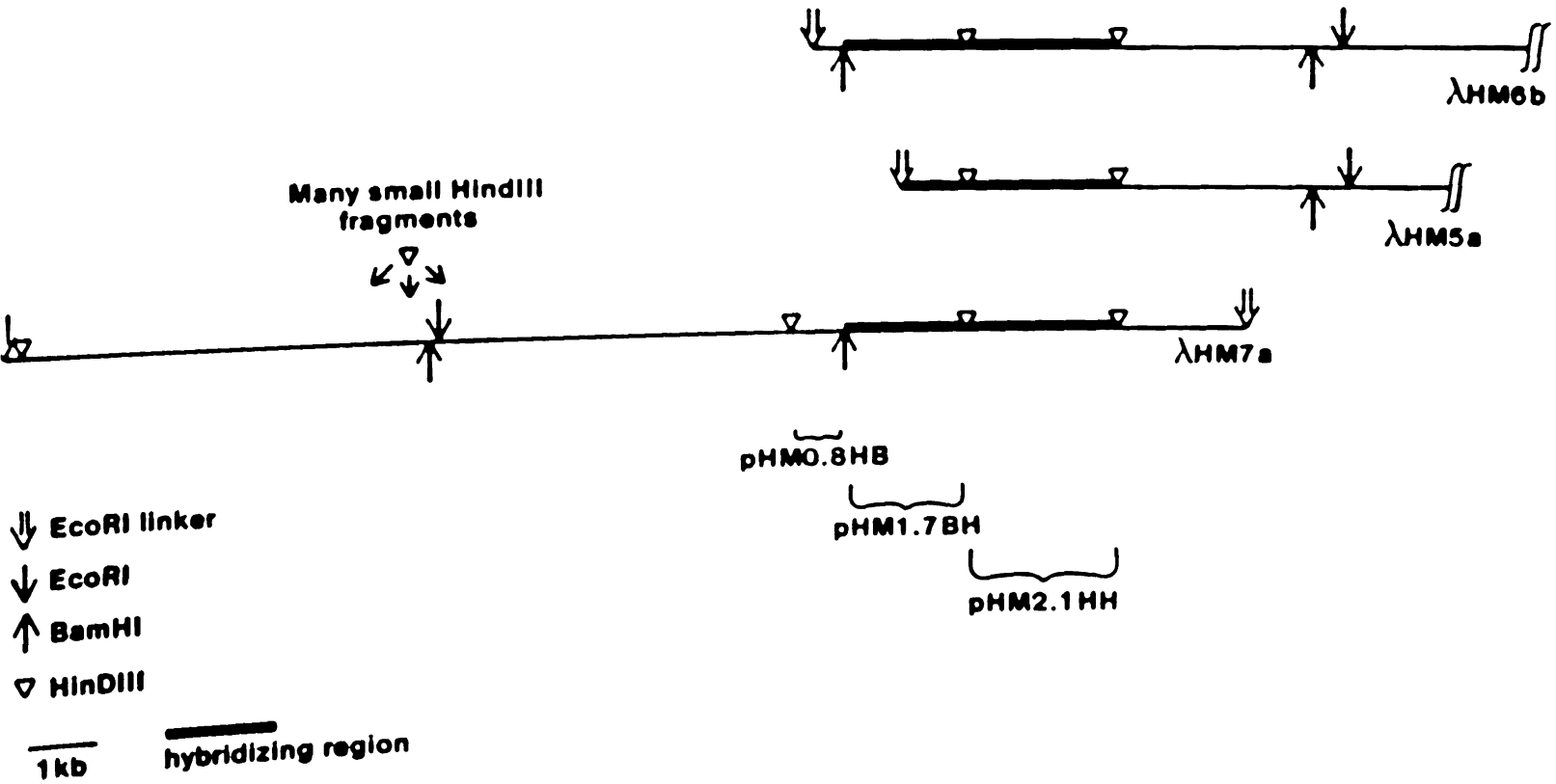

λHM6b
λHM5a
Many small HindIII fragments
λHM7a
pHM0.8HB
pHM1.7BH
pHM2.1HH
EcoRI linker
EcoRI
BamHI
HinDIII
1kb
hybridizing region

Figure 14

sites were used as labelling sites for sequence determinations by the chemical degradation method of Maxam and Gilbert. Figure 15 shows detailed restriction maps of pHM1.7BH and pHM2.1HH, emphasizing sites found in the cDNA and/or used for sequencing. The regions which were sequenced are indicated below the maps. Sequence determination showed that the chicken HMG-17 gene consists of 6 exons and 5 introns. The locations of the exons are indicated on the map in Figure 15.

Figure 16 shows the sequences of the 6 exons that are found in the HMG-17 cDNA. Exon 1 contains all of the 5' untranslated sequences, the initiation ATG, and 4 codons. Exons 2, 3, and 4 are quite small containing 45 bp, 30 bp, and 51 bp, respectively. Exon 5 is 96 bp long. Exon 6 contains 33 bp of coding sequence, the termination codon TAA, and the long 3' untranslated region of the gene. Because of the length of the 3' untranslated region, Exon 6 is much larger than the other exons. In the 3' untranslated region there are three nucleotide differences between the genomic sequence and the cDNA sequence. These were confirmed by sequencing both strands of the genomic clone. Whether these small differences are allelic in chickens or artifacts of cDNA synthesis was not investigated.

Figure 15. **Restriction maps of pHM1.7BH and pHM2.1HH. The locations of the 6 exons of the HMG-17 gene are shown. Arrows indicate the regions which were sequenced.**

Part of the HMG-17 promoter is on pHMO.8HB.

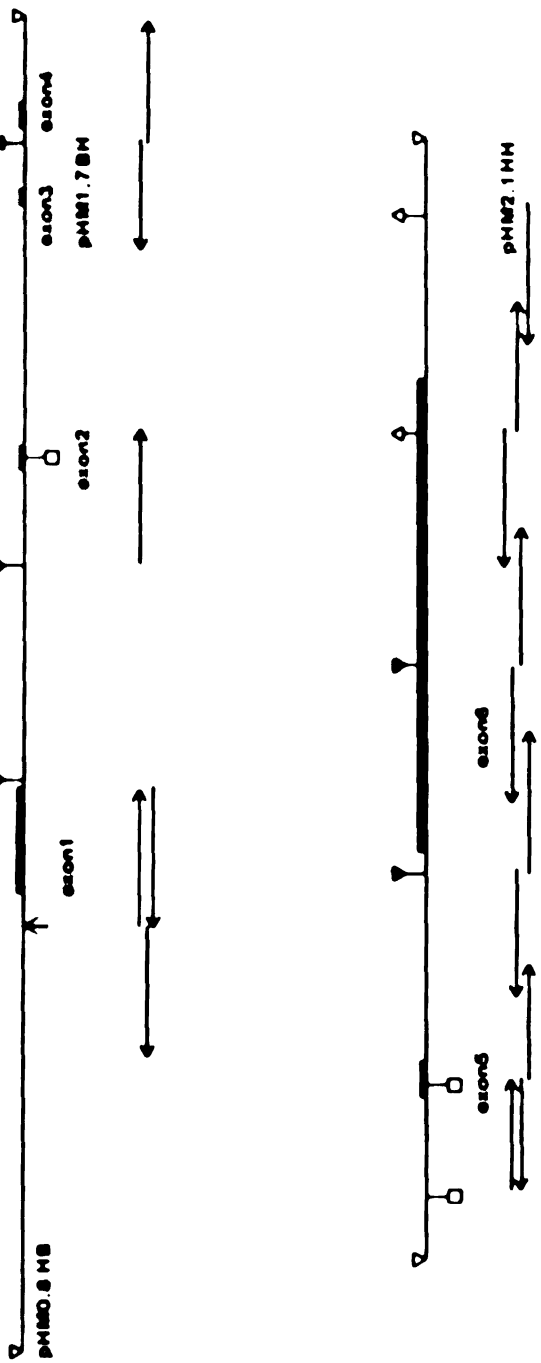

pHM0.8 HB
exon1
exon2
exon3
exon4
pHM1.7BH
exon5
exon6
pHM2.1HH
HinDIII
BamHI
ApaI
AvaI
AccI
StyI

Figure 15

Figure 16. **The exon-intron structure of HMG-17. The sequence of the cDNA pLG1a is shown; the locations of the 5 introns is indicated with arrows. The coding sequence is capitalized.**

HMG-17 cDNA insert

gaattccgca gccagcgcag cgaagccggc cgccagcccc gccgcgccgc cccgctctcc
ccctcggccc tccccccgctt ctcgccgcca ccgagcgagc ccggtgcccg cccccgcccg
ccccctccgc tcgctctctc cctcctcgca caacacacgc acgcgccgcc cggagctATG

Intron 1 Intron 2

CCGAAGAGAA AGGCTGAAGG AGATACCAAG GGCGATAAGG CCAAAGTTAA GGATGAGCCA

Intron 3

CAACGGAGAT CGGCAAGGTT ATCTGCTAAA CCTGCCCCTC CGAAGCCAGA GCCTAAACCT

Intron 4

AAAAAGGCAG CTCCAAAGAA GAGTGAGAAG GTGCCCAAGG GAAAGAAGGG GAAAGCTGAT

Intron 5

GCTGGCAAGG AGGGAAACAA CCCTGCAGAA AATGGAGATG CCAAAACAGA CCAGGCACAG

AAAGCCGAAG GTGCTGGTGA TGCCAAGTAA aatgtgtgaa tttttgataa ctgtgtactt

ctggtgactg tacagtttga aatactattt tttatcaagt tttataacaa tgcagaattt
tgttttactt tttttaagct atgttgttag cacacaaacc gctttgttgt tgtgttttga
ggggggggca gtgggacaaa cgtcacttaa tctgtttctt ggaacctaaa ttttaaaagt
ttaccccttc ccagtttttt agaaggactc ttcctaaatg gagcaggaag ggattccttc
gtgctgcaca cctcttccgt tttgtggacc gcatcagagt gaacggaagc tcccgagatg
cctgttgcca acttcagaac tgcagtttgc agtgccctct gcgtttcctt tcatgccctc
ccttttgcc tagagcctat cactccgaaa tacagcagac atggcatgtt gggactcacc
actctaaatg cattgtcagg tgatctgaac ttctggtgtc taatttggga tataatagct
ctaaaaggag ctgcatttcc tctttcatat tgtagatcta caaattaagg aatctgcagt
ttttaatttt tcctcgcaaa gctagggtag atttgtgaag agttgttaaa caacatgcta
aatgtgaaag tgtccgccct cactctaaac atttccctct acaagtatac aaaaatgaag
atttgtcggt tttatagcaa cctttatgtt tgggtagtcc atgaagggag gggagtttga
cagttgttgt aaaatgttgc agattgtagc ccatgtcctg cctaaattac catgattgtt
tatgaaaagt acctttaata aagctggata cggtttggct tggaaaaaaa aaaaaaaaaa
aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aacggaattc

Figure 16

Figure 17 shows the exon-intron boundaries of each of the introns. All of the introns start with the sequence GT, and all end with the sequence AG. This arrangement is seen at the ends of almost all introns in nuclear genes (4). In addition to the essentially invariant dinucleotides at the ends of introns, consensus sequences of preferred nucleotides around splice donor and splice acceptor sites have been seen. These structures are often:

```
...AG/gtr...intron...ttncag/N...
  (donor)            (acceptor)    .
```

The HMG-17 splice donors match the consensus sequence quite well except that the splice donor for intron 3 is CT/gtr, not AG/gtr. The splice acceptors do not match the consensus sequence well, except that 4 of the 5 introns end with the preferred cag. Intron 5 ends with tag, which is the next most common end.

The 3' end of the HMG-17 cDNA appears to faithfully represent the processed end of an HMG-17 mRNA. The cDNA ends with a 49 base polyadenylate tract; 27 bases upstream of this sequence is the canonical polyadenylation signal AATAAA (5). The genomic sequence of the 3' end of the HMG-17 gene shows that the polyadenylate tract is in fact added post-transcriptionally. The genomic sequence is identical to the cDNA sequence until the polyadenylate sequence of the cDNA. Figure 18 shows the 3' end of the HMG-17 gene, and

Figure 17. **HMG-17 exon-intron boundaries. Coding sequences are capitalized; intron sizes are indicated.**

Intron 1 CCGAAGAGAAGgtactggcacc...0.62kb...ccccgcgggcagGCTGAAGGAGAT

Intron 2 GTTAAGGATGAGgtaagaagttg...0.38kb...ctcttttggcagCCACAACGGAGA

Intron 3 AGGTTATCTGCTgtaagtattcc...0.24kb...tacactctacagAAACCTGCCCCT

Intron 4 CAGCTCCCAAAGgtgagttgctg...0.40kb...tgcgttcgacagAAGAGTGAGAAG

Intron 5 AAAACAGACCAGgtataacagca...0.40kb...ttttttcttttagGCACAGAAAGCC

Figure 17

Figure 18. **The 3' end of the HMG-17 gene. Sequences found in the cDNA pLG1a are capitalized; the polyadenylation signal is underlined; the polyadenylation site is indicated with an arrow.**

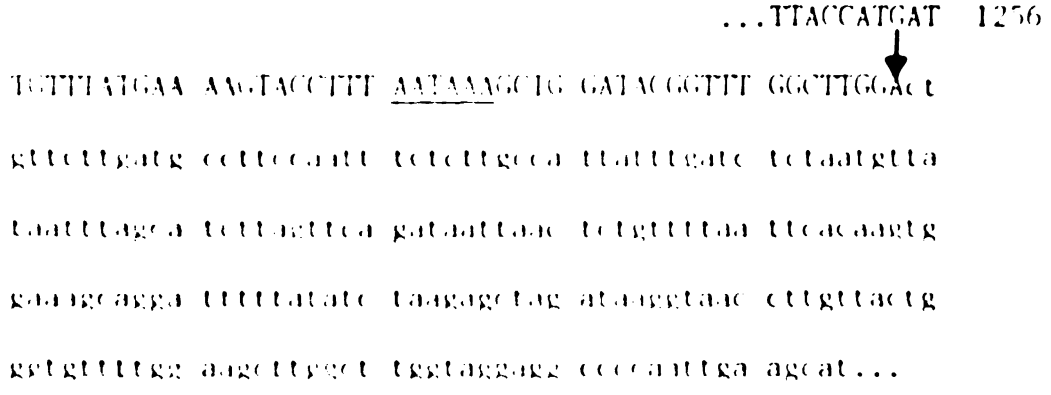

...TTACCATGAT 1256
TGTTTATGAA AAGTACCTTT AATAAAGCTG GATACGGTTT GGCTTGGact
gttcttgatg cctccattt tctcttgcca ttatttgatc tctaatgtta
taatttagca tcttacttca gatcattaac tctgttttaa ttcacaatg
gaaagcagga ttttatatc taaagctag atatgctaac cttgttactg
gctgttttga aagcttgcct tgctaggagg cccaattga agcat...

Figure 18

the polyadenylation site. The polyadenylation signal, AATAAA, is 27 bp upstream of the polyadenylation site. This spacing is similar to the spacing seen in other genes (5-7).

The HMG-17 promoter

The regions 5' of eukaryotic protein genes have been found to be important in the control of gene expression. These promoter regions often contain small sequence elements which have been well conserved (see Chapter 1). The most common of these is the TATAA element or Goldberg-Hogness box (8) which is similar to the prokaryotic TATA promoter and has a similar function (9). The TATAA element is recognized by part of the RNA polymerase II transcription complex, leading to accurate initiation of transcription at just one or a few nucleotides about 30 bp downstream of TATAA (10). In contrast to prokaryotic genes, not all eukaryotic protein genes have TATAA elements in their promoters. When the TATAA element is absent initiation of transcription usually can occur at a number of sites (11).

Another sequence element often found in promoters is CCAAT. This sequence is usually located 30 to 60 bp upstream of the TATAA element, but it may be found closer. Some promoters contain degenerate CCAAT elements, and some contain none at all (12).

The 5' regions of eukaryotic genes are often G+C-rich and contain G+C-rich elements which facilitate high levels

these elements is variable, but their effect seems to be additive and proportional to their proximity to the TATAA element (when one is present) (14). It is thought that these elements, like CCAAT and TATAA elements, are binding sites for proteins which participate in transcription. The first and best characterized of these proteins is the mammalian protein SP1 (15). This protein increases transcription by binding to a variety of G+C-rich elements which share the core consensus sequence GGCGGG. An avian homologue of SP1 has not been demonstrated, but G+C-rich sequences are common in the upstream regions of avian genes.

Figure 19 shows the promoter region of the chicken HMG-17 gene and exon 1 (in capitals). All of the common promoter elements are found in the HMG-17 promoter. Part of the HMG-17 promoter lies on pHM0.8BH, upstream of the subclones which contain the coding sequences. The HMG-17 promoter contains two TATAA elements. They are located 31 bp and 41 bp upstream of the first base represented in the cDNA clone pLG1a. This suggests that pLG1a is a nearly complete cDNA copy of an HMG-17 mRNA, perhaps missing a few transcribed nucleotides at the 5' end. Either TATAA element, or both, might function in this promoter. The site(s) of the start of HMG-17 mRNA transcription must be determined directly (see below), not by an examination of

Figure 19. **The HMG-17 promoter region. Exon 1 sequences are capitalized; common promoter elements are underlined; and mRNA initiation sites determined by S1 protection analysis are indicated with arrows.**

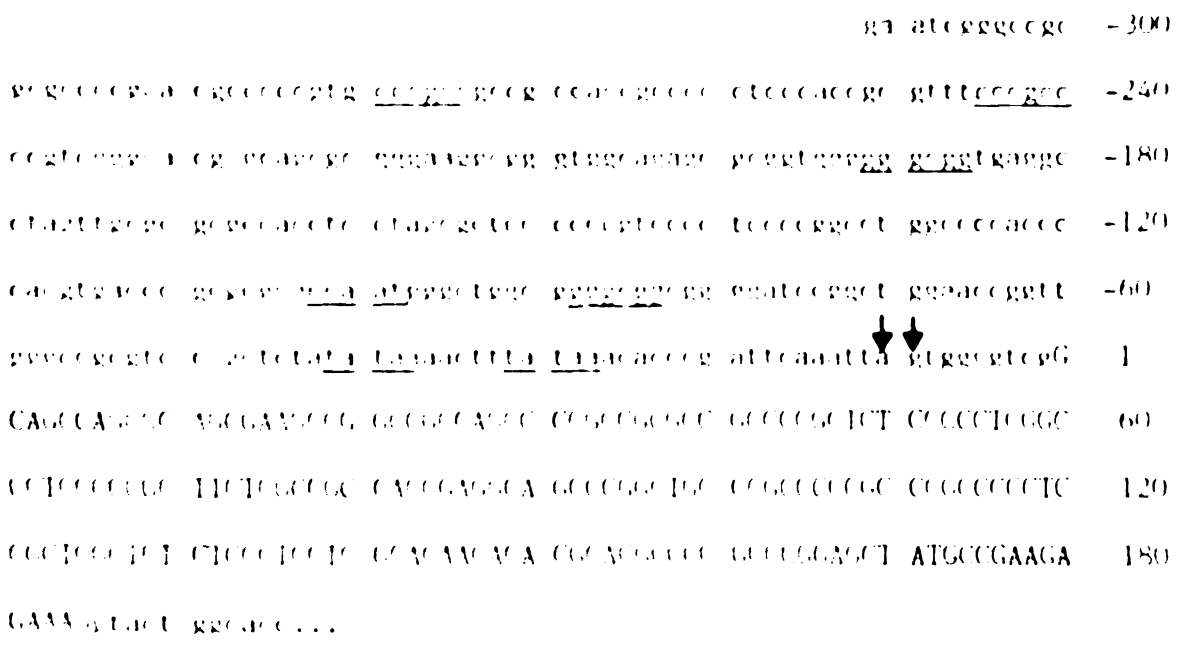


Figure 19

one cDNA clone. The chicken HMG-17 promoter also contains a CCAAT element 61 bp upstream of the 5' TATAA element. This spacing is similar to that seen in many eukaryotic genes.

The HMG-17 promoter has a G+C content of about 75%. Within this G+C-rich region are 4 SP1 binding sites. The 2 sites farthest upstream are on the non-coding strand of DNA, while the downstream elements, including one between the CCAAT and TATAA elements, are on the coding strand of the gene. This seemingly haphazard arrangement of SP1 binding sites is typical of this element.

S1 protection analysis

One method of determining where transcription of a gene starts involves S1 protection analysis. In this method, a radioactive probe is made which overlaps the end of the mRNA of a gene. This probe is hybridized to mRNA, then the hybrids are treated with S1 nuclease, which degrades single stranded nucleic acids, but not double stranded hybrid material. Measuring the portion of the probe which is protected from digestion by hybridization to mRNA can give an accurate indication of where transcription begins. Furthermore, when, as is usual, the hybridization probe is in excess of the corresponding mRNA, the intensity of the protected band is a good measure of the gene's mRNA level.

This method, as applied to the chicken HMG-17 gene, is

and +59 (Figure 19). This 118 bp Sau96a1 fragment overlaps the suspected transcriptional start site. After treatment with calf alkaline phosphatase, the fragment was labelled with γ^{32}P-ATP and polynucleotide kinase. After the labelled DNA was denatured, it was hybridized to mRNA isolated from several tissues. The resultant RNA:DNA hybrids were then treated with S1 nuclease and analyzed on a denaturing polyacrylamide gel. Figure 21 shows the results of this analysis. A small amount of untreated probe was run in lane P. The two labelled strands of DNA, each 118 nucleotides long, have slightly different mobilities because of their different base compositions. Some full length probe remains in each of the treated sample lanes. This material probably reannealed during the hybridization step and so became resistant to digestion by S1 nuclease. The amount of probe added to the RNA samples was 1000-fold more than the amount loaded in lane P, so the reannealed probe in lanes 1-12 is a small fraction of the original input.

When the 118 nucleotide probe was hybridized to total RNA isolated from embryonic brain, heart, skeletal muscle, or blood, a range of fragments centered on 70 to 71 nucleotides in length was protected from S1 nuclease digestion. This means that these tissues contain mRNAs homologous to the probe, and that these mRNAs start 70-71 bp upstream of the Sau96a1 site in exon 1. (It is formally

Figure 20. S1 protection analysis of HMG-17 mRNA.

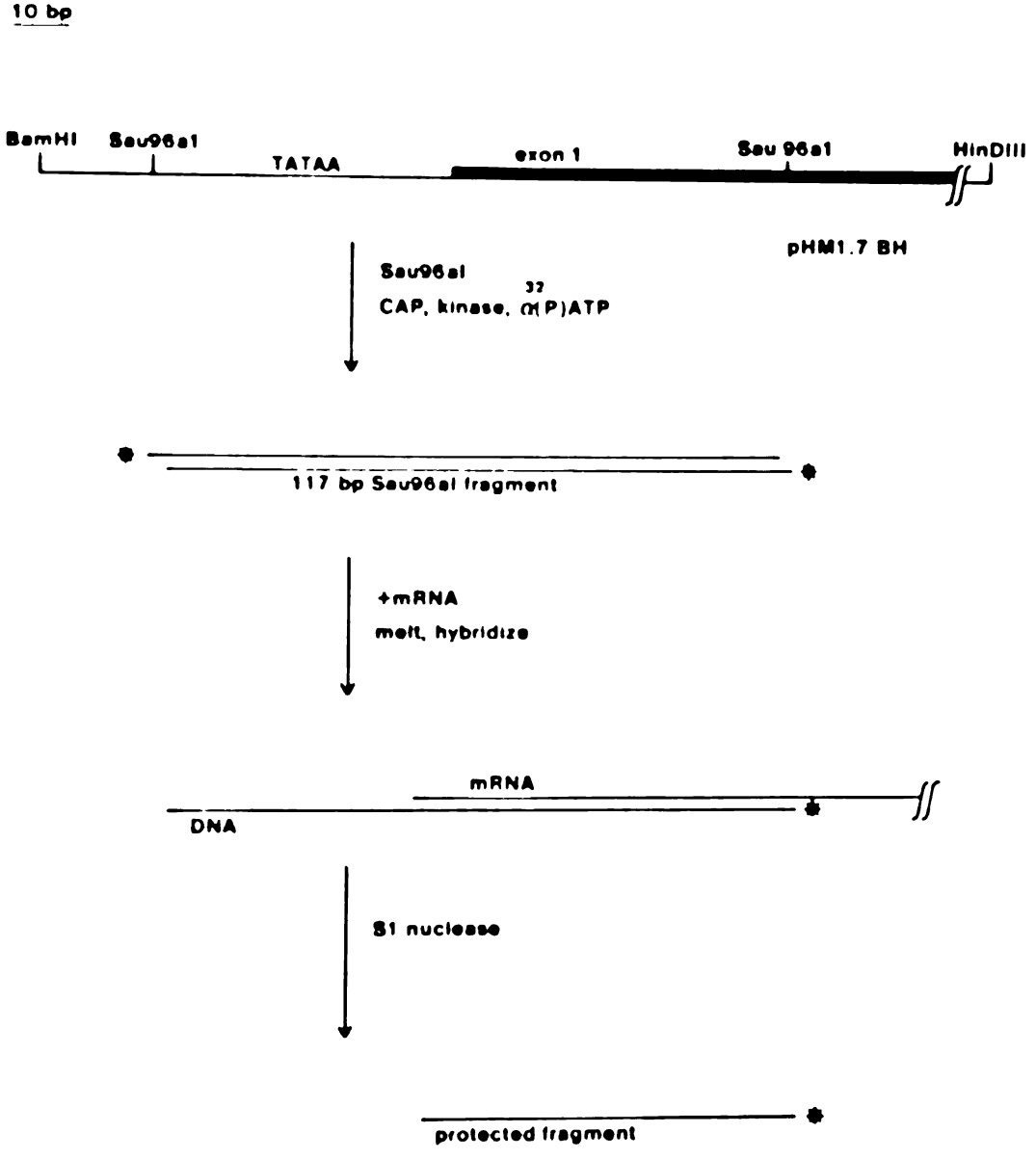
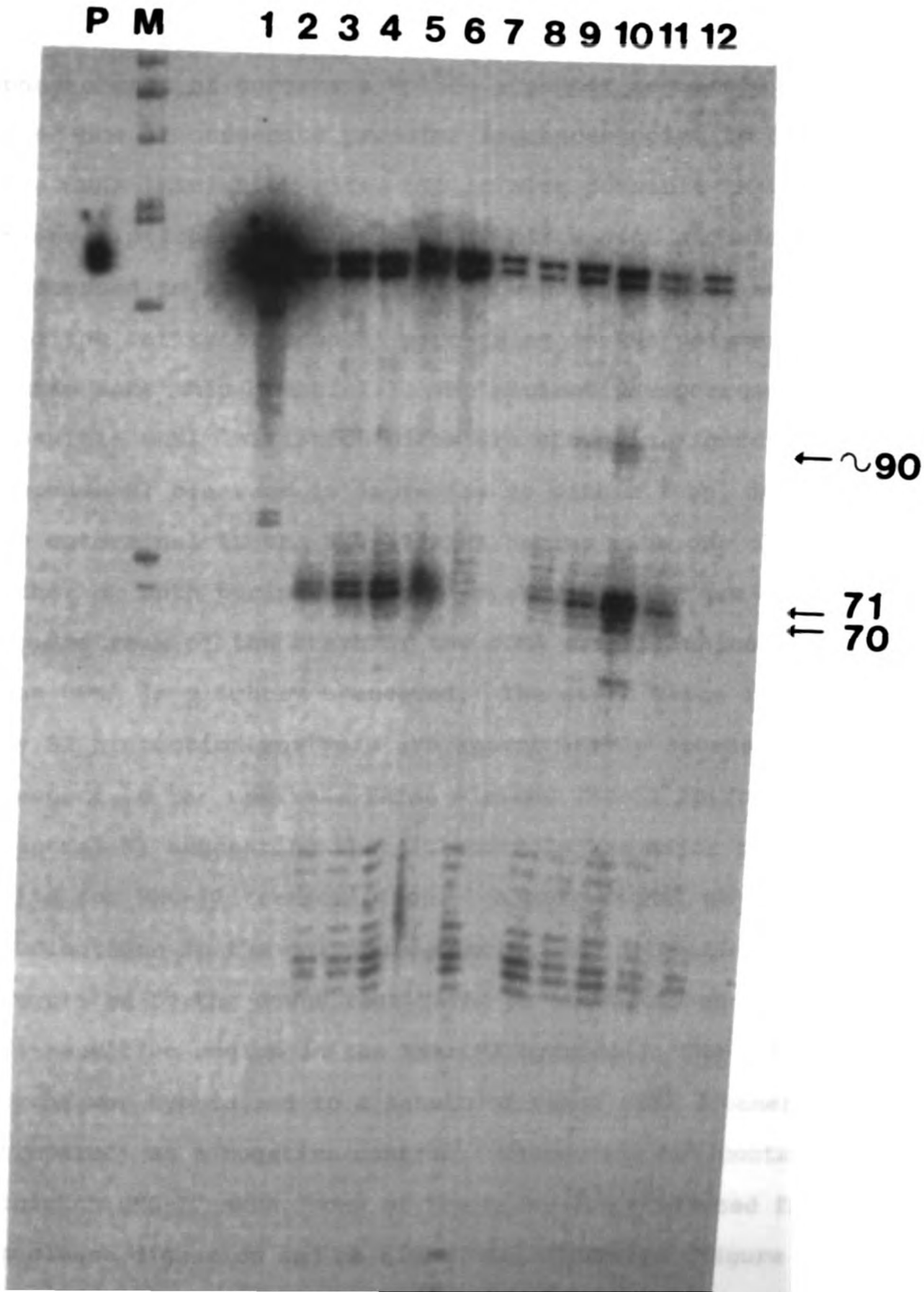

10 bp
BamHI
Sau96a1
TATAA
exon 1
Sau 96a1
HinDIII
pHM1.7 BH
Sau96a1
CAP, kinase, α(32P)ATP
117 bp Sau96a1 fragment
+mRNA
melt, hybridize
mRNA
DNA
S1 nuclease
protected fragment

Figure 20

Figure 21. Results of S1 protection analysis of HMG-17 mRNA from 15 day old chicken embryos. Lane P = probe only; M = pBR322 HinfI size marker. Experimental lanes: 1,7 = liver RNA
2,8 = brain RNA
3,9 = cardiac muscle RNA
4,10 = skeletal muscle
5,11 = blood RNA
6,12 = yeast RNA.

Samples in lanes 7-12 were treated with 2 times as much S1 nuclease as samples in lanes 1-6.
Signal sizes (in nucleotides) are indicated.


P M
1 2 3 4 5 6 7 8 9 10 11 12
~90
71
70

site, but its correspondence with the cDNA clone start site, the absence of consensus splice acceptor sequences, and the presence of consensus promoter sequences point to it being the mRNA initiation site. It is also possible that mRNA transcription began upstream of this region and was processed to mature mRNA, but consensus promoter sequences and the rarity of such 5' processing in RNA polymerase II mRNAs make this doubtful.) The nucleotides corresponding to possible mRNA initiation sites are shown in Figure 19. Because S1 cleavage is imprecise to within 2 bp, it cannot be determined if the HMG-17 mRNA begins with one or the other or both purine nucleotides shown. They are 9 and 10 bp upstream of the start of the cDNA clone, indicating that the cDNA is slightly truncated. The start sites indicated by S1 protection analysis are appropriately located with respect to the upstream TATAA element (30-31 bp from the central T) suggesting that it controls the major initiation site for HMG-17 transcription. (A weak signal at about 63 nucleotides in the gel could result from initiation regulated by the downstream TATAA or be due to an internal S1-sensitive region in the RNA:DNA hybrids.) The 118 bp probe was hybridized to a sample of yeast tRNA (commercially prepared) as a negative control. Since this RNA contains no chicken HMG-17 mRNA, none of the probe was protected from S1 nuclease digestion and no signal was generated (Figure 21,

S1 analysis also shows that the HMG-17 gene is not active in embryonic liver. Figure 21, lanes 1 and 7, shows that liver RNA prepared from 15 day old embryos does not generate the 70 and 71 nucleotide signals. The few faint larger bands seen in lane 1 are products of incomplete S1 digestion; when more S1 nuclease is used to digest the hybridization mixture these bands do not appear (lane 7).

A+T-rich regions of double stranded nucleic acids are somewhat susceptible to S1 digestion because of transient local melting of these regions. We suspect that the 2 TATAA elements on the reannealed 118 bp probe are slightly digested by high levels of S1 nuclease, generating a faint signal of about 90 nucleotides in some lanes.

Other Results

Landsman, Srikantha, and Bustin (17) reported the isolation and characterization of the chicken HMG-17 gene. They used a human HMG-17 cDNA clone to isolate genomic clones which contain the chicken gene. Their results agree with ours in every respect except one. They used the primer extension method to determine the mRNA initiation site. Their results predict that HMG-17 mRNA starts at the first A in the sequence TTCAAATTAGTGGGG, while we predict a start at the fourth A or its neighbor G (Figure 19). While their published results are not completely convincing (not shown), the tendency of S1 nuclease to digest A+T-rich hybrids may

S1 protocol, since they predict an mRNA with 6 As and Ts at the 5' end. The initiation site they predict is 24 bp 3' of the central T in the upstream TATAA. They have evidence, as do we, of minor utilization of the downstream TATAA. Their evidence predicts a minor mRNA species which starts in the G_4 sequence 8 bp downstream of the major initiation site. We have not attempted to resolve this minor discrepancy between our results.

Conclusion

Analysis of the chicken HMG-17 gene has shown that:

1. The gene is a single copy gene.
2. The gene contains 5 introns; the exon-intron boundaries are unremarkable.
3. The mRNA is polyadenylated 22 nucleotides downstream of the polyadenylation signal AAUAAA.
4. The HMG-17 promoter contains TATAA, CCAAT, and SP1 binding elements in a normal arrangement.
5. The gene is active in embryonic brain, heart, skeletal, muscle, and blood, but not in embryonic liver.

The chicken HMG-14a gene

To isolate the chicken HMG-14a gene, the chicken HMG-14a cDNA pLM3a was used as a radioactive hybridization probe in a screen of a λCharon4A library of the chicken genome as described in Chapter 2. Fourteen plaques tested positive in the initial hybridization screen of the library. When the 14 clones were isolated, replated and rescreened, 3 of them gave strong positive signals. These were named λYN1, λYN2, and λYN3. DNA was prepared from the purified clones and used for restriction mapping and Southern analysis. Figure 22 shows restriction maps of λYN1, λYN2, and λYN3. Southern blotting the mapping gels and hybridization of the HMG-14a cDNA probe to the blots allowed identification of the fragments which contain coding sequences of the gene, and facilitated alignment of the maps to show how these clones overlap.

Fragments of λYN1 and λYN3 which hybridize to the HMG-14a cDNA probe were subcloned into a plasmid vector (Bluescribe KS+). These subclones, pHM1.9HB, pHM1.3BH, pHM1.8HB, pHM1.0RH, and pHM1.5HP, were used for restriction mapping and sequence analysis. Identical pHM1.8HB subclones were isolated independently from both λYN1 and λYN3, confirming that these λclones contain overlapping sequences of the chicken genome.

Figure 22. **Restriction maps of recombinant λCharon4A chicken genomic clones which hybridize to the HMG-14a cDNA pLM3a. The clones are aligned to show how they overlap.**

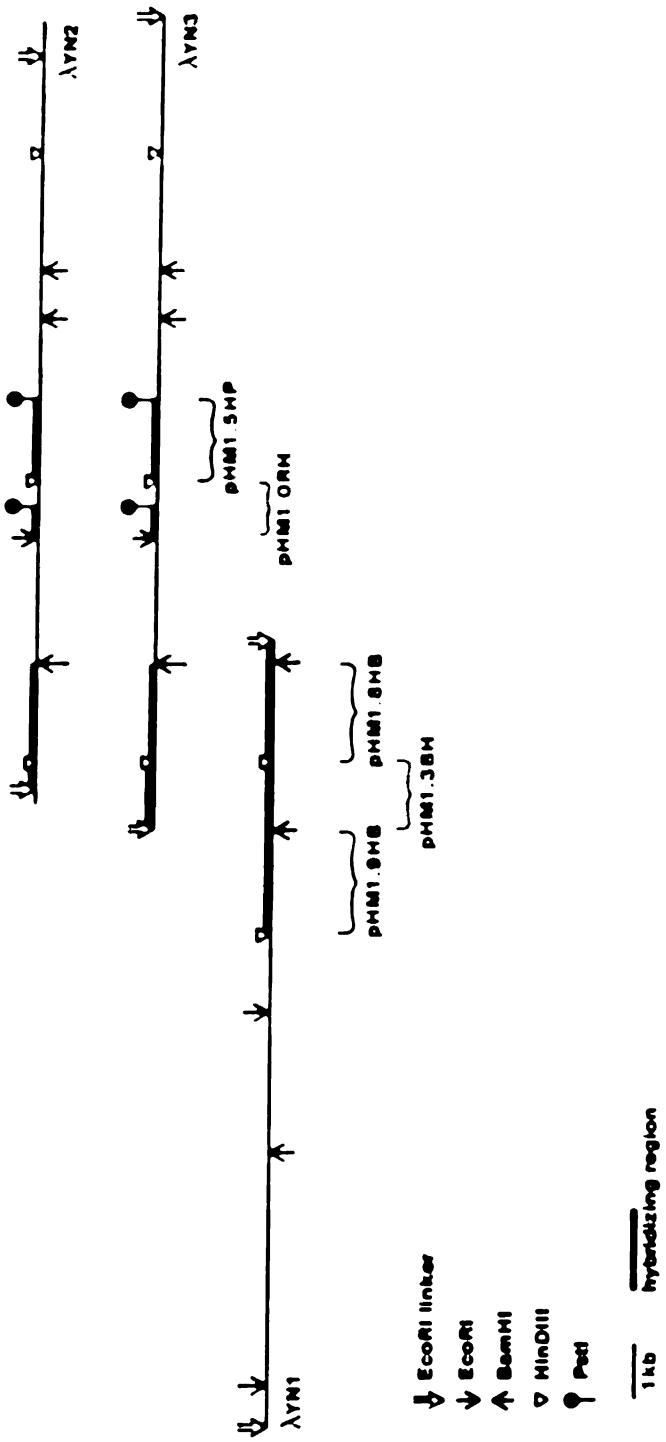

λYN2
λYN3
pHM1.5HP
pHM1.0RH
pHM1.8HB
pHM1.9HB
pHM1.3BH
λYN1
EcoRI linker
EcoRI
BamHI
HinDIII
PstI
1kb
hybridizing region

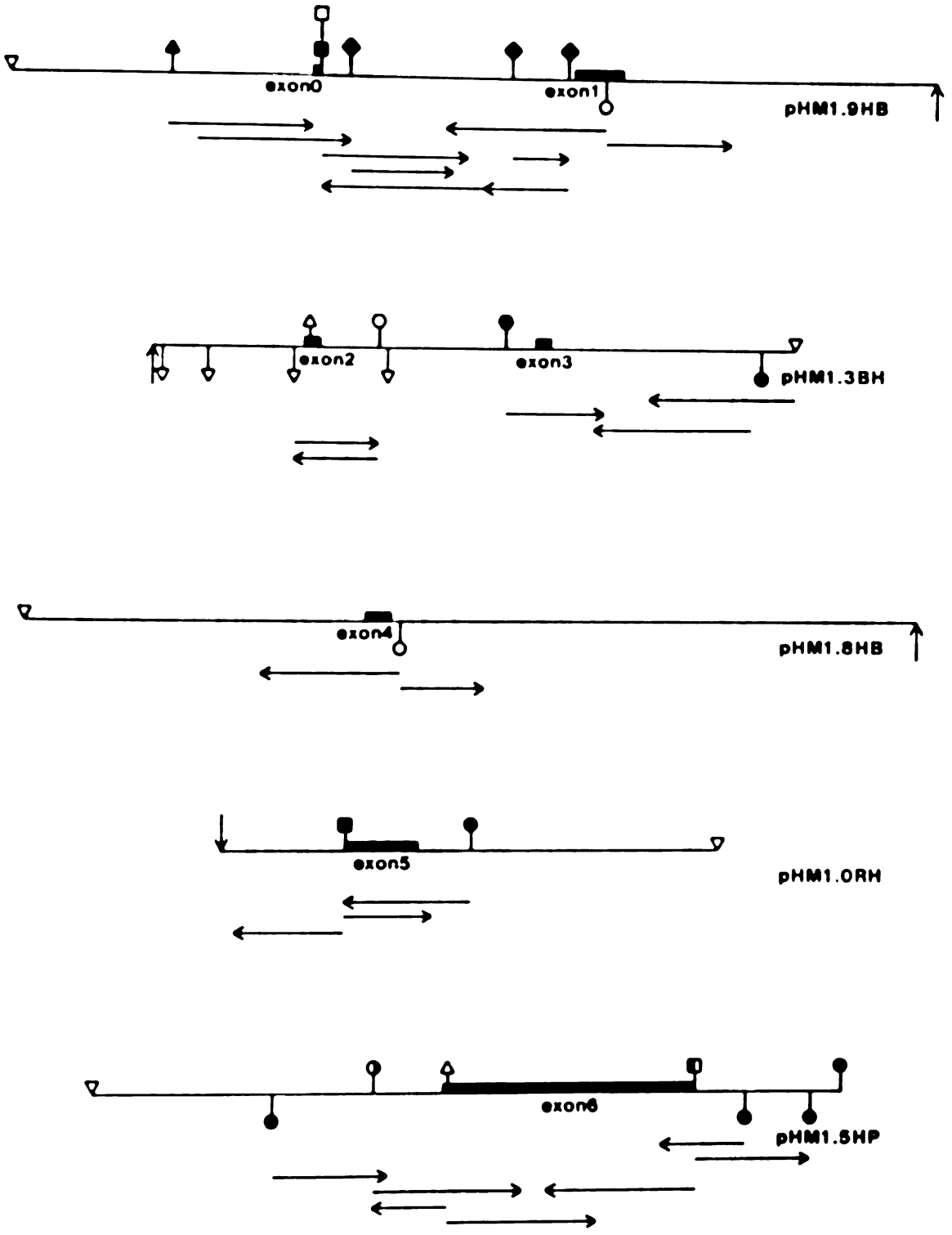
Figure 22

The region which hybridizes to the HMG-14a cDNA probe is about 10 kb. This is much larger than the chicken HMG-17 gene, which is less than 4 kb. Southern analysis of genomic DNA had shown that the chicken HMG-14a gene is a single copy gene, but might have weak homology to a few other sequences in the genome. It seemed possible that λYN1 and λYN3 might contain 2 homologous genes. Alternatively, the chicken HMG-14a gene might contain more intron sequences than the HMG-17 gene. Restriction mapping and sequence analysis eventually showed that there is only one large HMG-14a gene in this region.

Figure 23 shows detailed restriction maps of the subclones of λYN1 and λYN3. As in the analysis of the HMG-17 gene, restriction sites present in the cDNA clone were suggestive of the locations of exons. In some cases these sites were used as labelling sites for sequence determination by the chemical degradation method. The regions which were sequenced are indicated below the maps in Figure 24.

Sequence analysis showed that the chicken HMG-14a gene consists of 7 exons and 6 introns. For reasons discussed below the exons are numbered exon 0 through exon 6. Figure 24 shows the sequences of the 7 exons that are found in the HMG-14a cDNA pLM3a. Exon 0 is very small, consisting of

Figure 23. **Restriction maps of plasmid subclones which contain HMG-14a exons. The exons are indicated with heavy bars. Arrows indicate regions which were sequenced to locate the exons.**


exon0
exon1
pHM1.9HB
exon2
exon3
pHM1.3BH
exon4
pHM1.8HB
exon5
pHM1.0RH
exon6
pHM1.5HP
1cm = 100bp
HinDIII
Asp718
BamHI
TaqI
PstI
AccI
ApaI
EcoO109
Sau3a1
SstI
OxaNI
HincII
PvuII
HinfI
DdeI

untranslated region, the initiation ATG, and 4 codons make up exon 1. This is similar to exon 1 of HMG-17. Exons 2, 3, and 4 of HMG-14a are quite small, as they are in HMG-17. In HMG-14a they consist of 30 bp, 30 bp, and 51 bp, respectively. Exon 5 is somewhat larger at 144 bp. Since exon 6 contains the long 3' untranslated region of the gene, it is much larger than any of the other exons.

The relative sizes of the exons of HMG-14a and HMG-17 are similar. In fact, 3 of the exons contain coding regions for virtually identical regions of the homologous proteins. Figure 25 shows a comparison of the codons present in exon 1, exon 3, and exon 4 of the two chicken HMG genes. The regions of the protein specified by exon 1 of each gene have identical amino acid sequences. The same is true of exon 4 of each gene. The portions of the two proteins specified by each exon 3 differ by just one amino acid. Conservation of exon structure and intron location is seen in many other gene families, for example, between the α- and β-globin genes (4) and even where the predicted evolutionary separation of the homologues is as ancient as the 2 characterized chicken histone H3.3 genes (17).

Figure 24. **The exon-intron structure of HMG-14a. The sequence of the cDNA pLM3a is shown; the locations of the 6 introns is indicated with arrows. The coding sequence is capitalized. Sequences which are complementary to the synthetic oligonucleotides JD-20 and JD-21 are underlined.**

HMG-14 cDNA insert

Intron 0

gaattccgtc ccc<u>ttcctca ggacgctcga aaa</u>cagttt<u>c tcggcggttc ccttcctatt</u>

JD-21 JD-20

ttttacacct ctcccgatct ctctatttgc agtcaactat taaggtgcaa ctATGCCCAA

Intron 1 Intron 2

AAGAAAGGCT CCAGCTGAAG GCGAGGCGAA GGAGGAGCCA AAGAGAAAGT CGGCCAGACT

Intron 3 Intron 4

ATCTGCTAAA CCTGCTCCGC CTAAACCGGA GCCAAAGCCC AAAAAGGCAG CACCTAAGAA

AGAAAAGGCA GCAAACGATA AAAAGGAAGA CAAAAAGGCA GCAACAAAAG GGAAGAAAGG

AGCCAAAGGC AAAGACGAAA CTAAACAAGA GGATGCAAAA GAAGAAAACC ACTCTGAAAA

Intron 5

TGGAGATACC AAAACTAATG AGGCACCAGC TGCTGAAGCA TCTGATGATA AGGAAGCCAA

GTCCGAGTAA tgttaaccct gccctatatc tccatcattt ggtatccgta cctccatgct
gtattgttaa cagagaggaa tatttttatc aactatttta taaatgcagg ttttttagc
atgaatttaa ttatggaaca tcttcatctc ggttacttgg gaattaaatc cctaacaaac
aaaacaaaac aaaacaaaaa aaaatcattg ttttaaattt gtgattgtaa tagtttgtat
ggtacatgga aagaataagt ggtggtagct tttgacttct gtcagtgtgt ccctttttgt
gtaagtcatg cttacagact tcagatttta attttaccct tgtatgtgtt gtatggtttc
ttaaagtggg gaggtctcaa aacagataac tgtgttaaac attccagtgg ttctgtgggt
tgcttttata aagaaggtga gctattttca tgaaaaaaaa aaaaaaaaaa aacggaattc

Figure 24

Figure 25. **Conserved exons of HMG-14a and HMG-17. These portions of the proteins differ at only one amino acid.**

Exon 1

HMG-14: ATG CCC AAA AGA AAG

HMG-17: ATG CCG AAG AGA AAG

Exon 3

lys

HMG-14: CCA AAG AGA AAG TCG GCC AGA CTA TCT GCT

HMG-17: CCA CAA CGG AGA TCG GCA AGG TTA TCT GCT

gln

Exon 4

HMG-14: AAA CCT GCT CCG CCT AAA CCG GAG CCA AAG CCC AAA AAG GCA CCA CCT AAG

HMG-17: AAA CCT GCC CCT CCG AAG CCA GAG CCT AAA CCT AAA AAG GCA GCT CCA AAG

Figure 25

Figure 26 shows the exon-intron boundaries of each of the introns. The introns begin with GT and end with AG, as do virtually all introns of nuclear genes. All of the splice donors match the consensus sequence AG/gtr except that of intron 3, where the splice donor is CT/gtr. The splice acceptors do not match the consensus sequence ttncag/N well, except that 5 of the 6 introns end with the preferred cag. Intron 5 ends with tag, which is the next most common end (4).

Exon 0

To show that exon 0 is not an artifact of cDNA cloning and to determine the structures of the 5' end of the HMG-14a transcript several types of experiments were done:

1. Exon 0 and intron 0 were characterized by further genomic DNA sequencing.
2. Additional HMG-14a cDNA clones were characterized.
3. S1 protection analysis and primer extension were used to characterize HMG-14a mRNA from several tissues.

Exon 0 contains all but the last nucleotide of an OxaNI site, CCTNAGG. If exon 0 abuts a normal intron (which would start with a G), this OxaNI site will be found in the genome. Restriction mapping showed that there is only one OxaNI site within 2.5 kb of exon 1. Sequence analysis around this OxaNI site confirmed that this site marks the location of exon 0 in the genome.

Figure 27 shows the sequence of the chicken genome around exon 0, including all of intron 0 and part of exon 1. Intron 0 is 643 bp long, from -482 through +161. (Nucleotide numbering will be explained below.) The splice donor and acceptor sites of intron 0 match the consensus sequences as well as do the other HMG-14a introns (Figure 26). Since exon 0, as represented in pLM3a, is very small, it seemed possible that pLM3a is a truncated cDNA which is missing sequences representative of the 5' end of HMG-14a mRNA. One direct way to determine this would be to examine other HMG-14a cDNA clones.

The cDNA library which was originally screened for HMG clones was constructed in the vector λgt11. cDNA copies of mRNA are cloned into this vector with the use of EcoRI linkers (18). After recombinant clones are identified and purified, the cDNA inserts are excised with EcoRI and

Figure 26. **HMG-14a exon-intron boundaries. Coding sequences are capitalized; intron sizes are indicated.**

Intron 0 CCCCTTCCTCAGgtaccgtcccg...0.54kb...gcaatcccgcagGACGCTCGAAAA

Intron 1 CCCAAAACAAAGgtagatgggggc...0.88kb...cttctcttgcagGCTCCAGCTGAA

Intron 2 GCGAAGGAGGAGgtaggatttcc...0.25kb...ccttattgccagCCAAAGAGAAGG

Intron 3 AGACTATCTGCTgtgagtacgtg...1.08kb...tgttttgtacagAAACCTGTCCGC

Intron 4 CAGCACCTTAAGgtaagtgttca...3.40kb...tttcttaaacagAAGAAAAGGCAG

Intron 5 AAAACTAATGAGgtaccttagct...1.48kb...acacttttatagGCACCAGCTGCT

Figure 26

Figure 27. **The HMG-14a promoter region. Exon 0 (as determined by primer extension) and exon 1 are capitalized. Only a portion of exon 0 is found in pLM3b; that part is underlined. The promoter contains no TATAA or CCAAT elements; SP1 binding sites are underlined.**

gca taacagtgtg ccacgagccc tgctggaacg -840
cgtagttagg acaaaacgga agcactggtc ccgccctgga cctcagcaca gggtctacct -780
cacctcaccc cgagtgcccg ctccaaacag cacctcgggc cgctgggtg agacggagcc -720
gagcgaacgg cgcctaccag ctgtagcccg ggccgacagc ccccgccccgc cgctgccccgt -660
cccggagtca ctcgggctgc gcggctgcac agctccgtgc cgcctcctcc tGCCCCCTGC -600
CCCGCACCTT CCTCCCCCCG CCGACACCAA CACCCCACCC ACTCCCCCCC CCGACCTCGG -540
CTCCCCTGCC GCGACAAACC CACCCCTCCC AAGCCCCCCC CGAAGCCCCT TCCTCAGgta -480
ccgtcccgct cccgcggacc ctcccccgccg ccaccgctcc cccccccccg ggcccgaggc -420
gcgagcccc gcgcatgtcg gcggcccccgc gcaggacgga gcggccgcca tgttgtcgtc -360
ctgcggcctg cgctaacttt gggtaagttt ctttcccagc ctctgggggg acgcgaggaa -300
ggagtagacg gggtcgagag gggggcgctc ggcggccgcg ggcagcaccg cacggcgagc -240
cgttccgaga gcggcagtga cgagggccg ccgcctcgca ccgacgcctc ccgcccctcca -180
cgagccgccg gggggggggg ggcgccgtat tcaccgcgcg cgttagcgcc tccgtgcgca -120
gacagcgtaa gcgggccgg ccctccgtgt gcgagcgtct ctcgcgctac gcctcgcccc -60
gccccccgcc ccccccccgc cccgccgcgc ttcccgcgcg cccattggct gcggcggccg 1
tcactctgcg gctgggccca tctcgcgctc cccattcatg cagtttggtt gcgtgttggt 61
gtattttact gtttgcaata aagaggggg gatttttttt ttgttcatct cttttttttt 121
ttctttttt cttctttttt ttttttttgc aatcccgcag GACGCTCGAA AACAGTTTCT 181
CCGCCGTTCC...

Figure 27

subcloned into a plasmid vector. This facilitates restriction mapping and sequence analysis since plasmid vectors are much smaller than λ vectors. Some cDNA inserts cannot be excised from the λgt11 vector by EcoRI. This is usually because one end of the cDNA did not receive an EcoRI linker, but was cloned by a blunt end ligation directly to vector DNA or other cDNA. Anomolously cloned inserts can be excised with restriction enzymes which cut the vector away from the cloning site, but the excised fragment will necessarily contain some λgt11 sequences.

Three such anomolous cDNA clones which gave strong positive signals in the original screening of the λgt11 library (Appendix) had been stored, uncharacterized. To clarify the nature of the 5' end of the HMG-14a gene, DNA from these clones was purified and analyzed. Restriction mapping of the purified recombinant λDNA showed that the 3 uncharacterized clones are identical. The cDNA insert, along with some vector DNA, was excised from one of them and subcloned into a plasmid vector (Bluescribe SK+). This subclone is pLM2a; its map is shown in Figure 28.

Restriction mapping showed that pLM2a is the same as pLM3a throughout most of the 3' part of the cDNA, but differs in the 5' region. In particular, pLM2a does not have the OxaNI site which characterizes exon 0. The sequence of the 5' end of pLM2a was determined by the dideoxy chain termination method, using a primer (SK) which

is complementary to the plasmid vector, and a primer (JD-20, see Figure 23) which is complementary to part of exon 1. This sequencing strategy is shown in Figure 28.

Sequence determination confirmed what restriction mapping had suggested: exon 0 is not found in pLM2a. As shown in Figure 29, pLM2a contains only exon 1 at its 5' end; but this exon 1 is 161 bp longer than exon 1 of pLM3a. Exon 1 of pLM3a contains the splice acceptor, at nt +162, to which exon 0 can be spliced (Figure 27). (The first nucleotide of pLM2a is arbitarily numbered +1.) The structures of pLM3a and pLM2a, when compared, imply that the initiation of HMG-14a mRNA synthesis might occur at 2 sites. Only mRNA which initiates at the upstream site will contain exon 0 for splicing onto exon 1(Figure 30). (Recent results obtained by other members of the lab involved characterization of 6 more independent HMG-14a cDNA clones. All were of the unspliced pLM2a type, i.e., lacking the 0 exon. All of these were isolated from a commercial total chicken embryo library and had shorter 5' ends than pLM2a. This suggests that this library contains a high level of 5'-truncated cDNAs, an observation confirmed with other clone types. Therefore, this clone set is not of use in trying to infer start sites of transcription of the pLM2a form of HMG-14a.)

To determine if exon 0-exon 1 splicing occurs in a significant portion of the HMG-14a mRNA population, and is

Figure 28. **Restriction map of pLM2a. The cDNA portion is joined to the vector portion without benefit of an EcoRI linker. Regions sequenced from the named primers are indicated with arrows.**

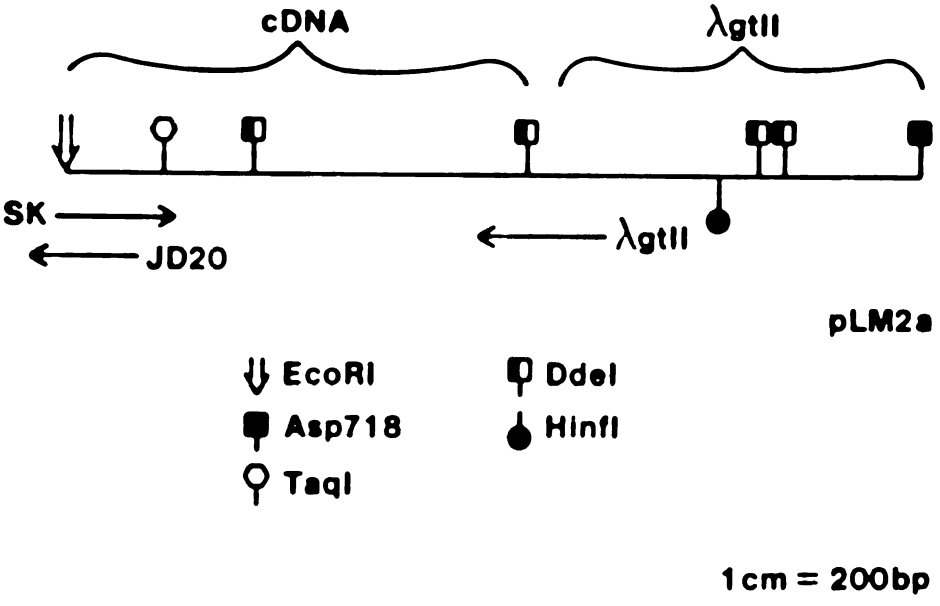

cDNA
λgtII
SK
JD20
λgtII
pLM2a
EcoRI
Asp718
TaqI
DdeI
HinfI
1cm = 200bp

Figure 28

Figure 29. **The HMG-14a promoter region. Sequences which found in pLM2a are capitalized. Exon 0 is not represented in pLM2a.**

gca taacatgtgt ccacgagccc tgctggaacg -840
cgtagttacg acaaaacgga agcactggtc ccgccccgga cctcagcaca gggtctacct -780
cacctcaccc cgaggccccg ctccaaacag cacctcgggc cgctggggtg agacggagcc -720
gagcgaacgg cgcctaccag ctgtagccgg ggccgacagc ccccgccccgc cgctgccccgt -660
cccggaggca ctcggctgc gcggctgcac agctccgtgc cgcctcctcc tgccccctgc -600
ccggcacctt cctgccccgcg ccgacaccaa caccgcacgc actccgcgcc ccgacctcgg -540
ctccccgtggc gggacaaagc caccgctcgc aaggccgccc cgaagccccct tcctcaggta -480
ccgtccccgct cccgcggacc ctccccgccg ccaccgctcc cccccccccg ggccccgaggc -420
gggagggcccc gcgcatggcg gcggcccccgc gcaggacgga gcggccgcca tgttgtcgtc -360
ctgccggcctg cgctaacttt gggaaggttt ctttcccagc ctctgggggg acgcgaggaa -300
ggaggagacg gggggggagag gggggggctc ggcggccgcg ggcagcaccg cacggcgagc -240
cgttccgaga gcggcagtga cgagggggccg ccgcctcgca ccgacgcctc ccgccctcca -180
cgagccgccg gggggggggg ggcgccgtat tcaccgcgcg cgttaggcgc tccgtgcgca -120
gacagcgtaa gcggtgccgg ccctccgtgt gcgagcgtct ctcgcgctac gcctcgcccc -60
gccccccgcc cccgcccccgc ccgccgcgc ttcccgcgcg cccattggct gcggcggccG 1
TCACTCTGCG GCTGGGCCCA TCTCGCGCTC CCCATTCATG CAGTTTGGTT GCGTGTTGGT 61
GTATTTTACT GTTTGCAATA AAGAGGGGGG GATTTTTTTT TTGTTCATCT CTTTTTTTTT 121
TTCTTTTTTT CTTCTTTTTT TTTTTTTGGC AATCCCGCAG GACGCTCGAA AACAGTTTCT 181
CGGCGGTTCC...

Figure 29

Figure 30. **Multiple splicing patterns of HMG-14a mRNA as represented by two cDNA clones. Exon 0 is found in only a fraction of the mRNA; it is spliced to an acceptor within exon 1.**

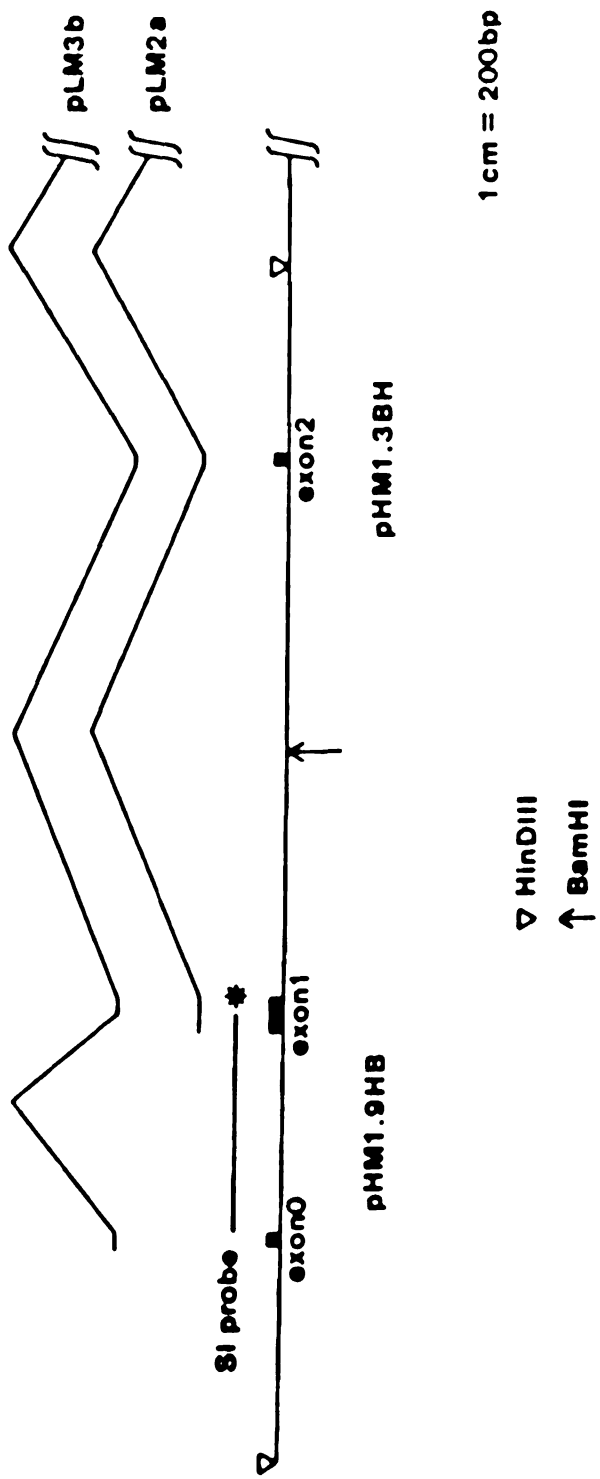

pLM3b
pLM2a
1cm = 200bp
exon2
pHM1.3BH
exon1
exon0
pHM1.9HB
SI probe
HinDIII
BamHI

Figure 30

not a rare aberrant event immortalized by cDNA cloning, S1 protection analysis was performed. The design of this experiment is shown in Figure 31. A DNA fragment which contains intron 0 and part of exon 1 was isolated. The strand which is complementary to HMG-14a mRNA was radioactively labelled at its 5' end and hybridized to chicken RNA from several tissues. After treatment of the hybridization mixture with S1 nuclease, only that portion of the probe complementary to exon 1 will remain intact. This radioactive fragment can be visualized and measured on an autoradiograph of a denaturing polyacrylamide gel. Figure 32 shows the results of this analysis. Some undigested probe remained in the experimental samples; a sample of untreated probe was run on the gel to identify these background signals. Chicken RNA (lanes 1-4), but not yeast RNA (lane 5), protects an 88 nucleotide fragment of the probe. This is the part of the probe which is complementary to the spliced exon 1 of pLM3a. These results show that the exon 0-exon 1 splice that is represented in pLM3a occurs in a significant fraction of HMG-14a mRNA, and is not a rare aberrant event.

Hybridization of the probe to HMG-14a mRNA containing unspliced exon 1, as represented in pLM2a, should result in the protection of a radioctive fragment of 249 nucleotides. No strong signal of this size is seen in Figure 32. Two explanations for this are possible.

Figure 31. **S1 protection analysis of HMG-14a mRNA.**

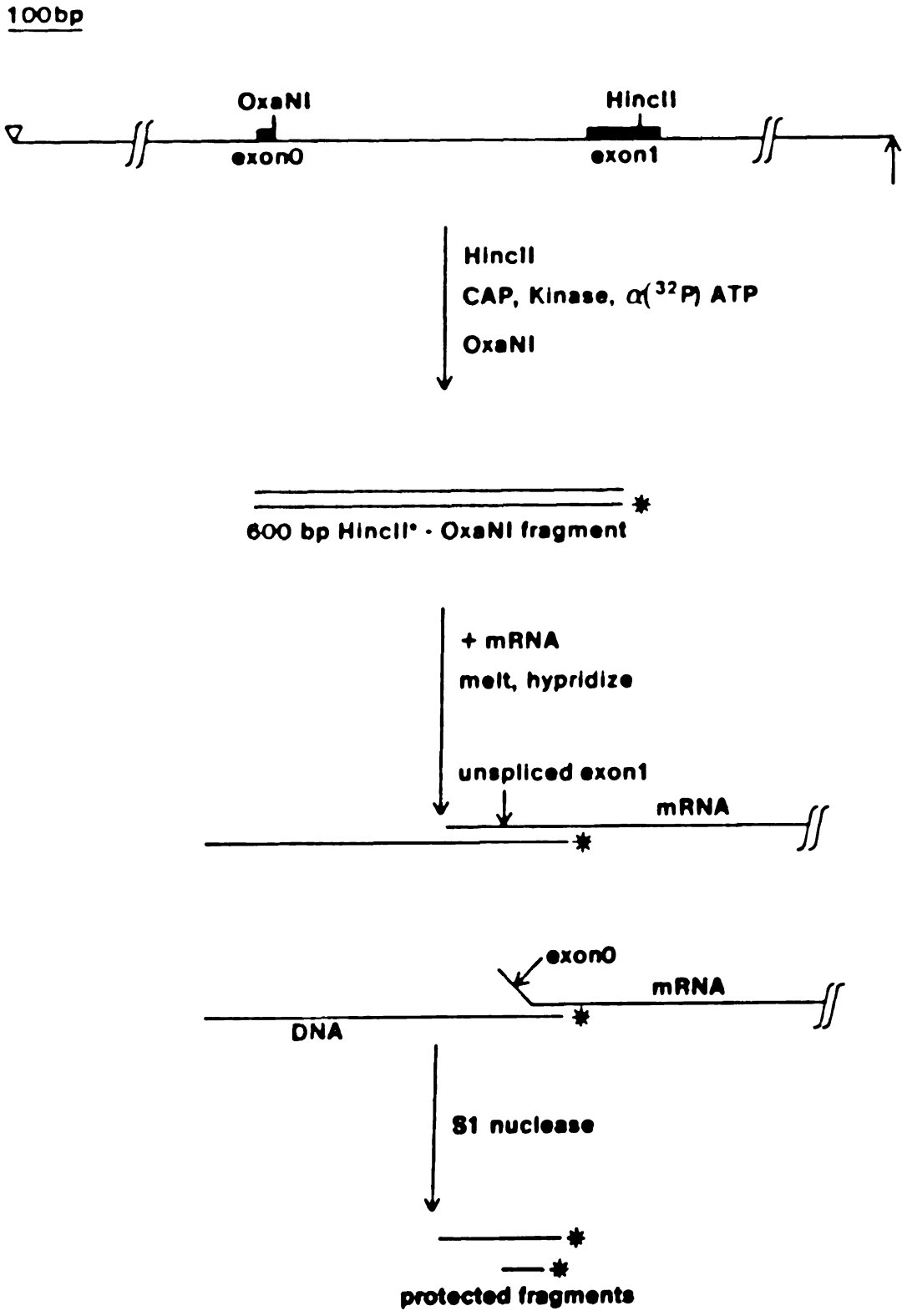
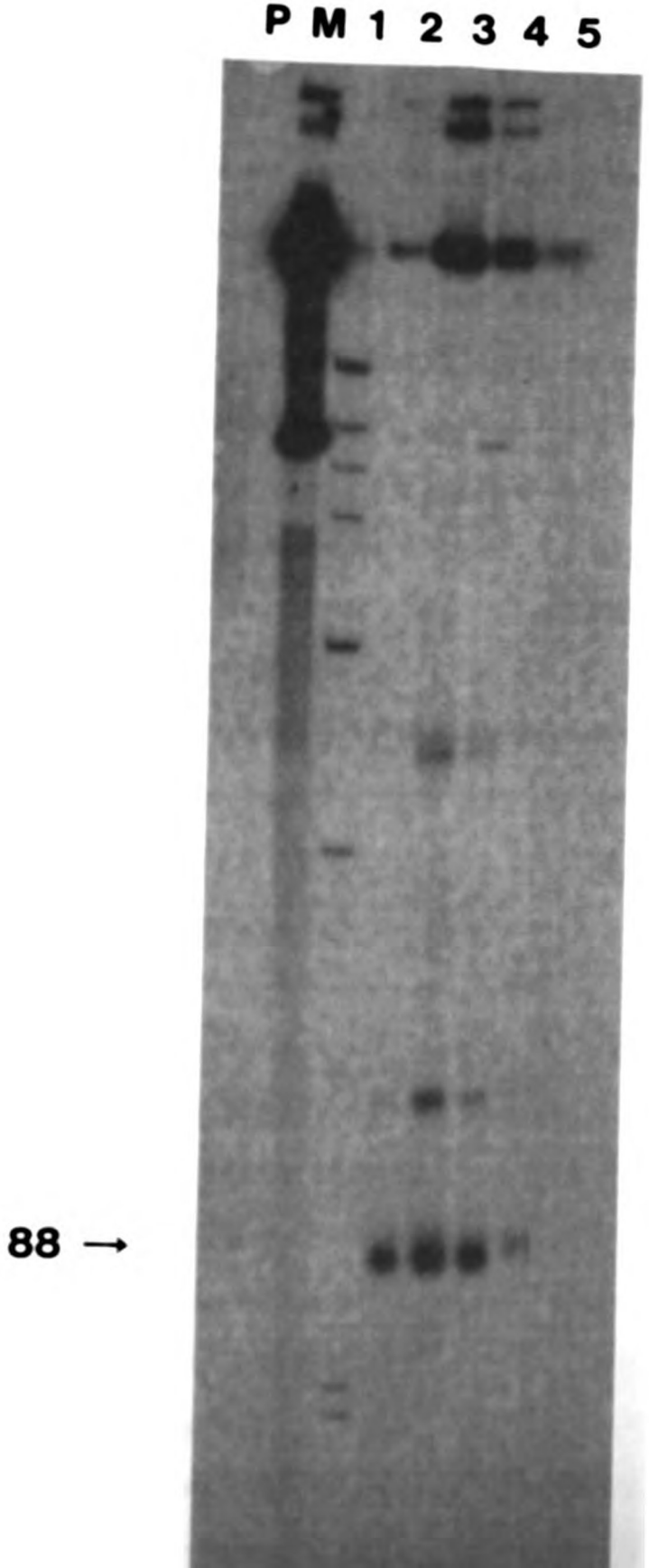

100bp
OxaNI
HincII
exon0
exon1
HincII
CAP, Kinase, α(32P) ATP
OxaNI
600 bp HincII* - OxaNI fragment
+ mRNA
melt, hypridize
unspliced exon1
mRNA
exon0
mRNA
DNA
S1 nuclease
protected fragments

Figure 31

Figure 32. Results of S1 protection analysis of HMG-14a mRNA. Lane P = probe only. Lane M = pBR322 HinfI digest marker.

Experimental lanes: 1 = liver RNA
2 = brain RNA
3 = cardiac muscle RNA
4 = skeletal muscle RNA
5 = yeast RNA.

The 88 nucleotide signal from mRNA containing the exon 0-exon 1 splice is marked. Larger signals come from unspliced mRNA.


P M 1 2 3 4 5
88 →

First, pLM2a may represent a rare form of HMG-14a mRNA which is not detectable by S1 protection analysis. This is unlikely from the recent cDNA clone results. Another explanation for the absence of 249 nucleotide signal from unspliced exon 1 is that hybrids of the probe and exon 1 might be sensitive to S1 nuclease despite being double stranded. Regions rich in A:T base pairs are sensitive to single stranded nuclease, probably because transient local melting of the RNA:DNA hybrid occurs in these regions. Inspection of the sequence of exon 1 (Figure 29) shows that it contains a stretch of 56 bp which is 89% A:T base pairs. This region of probe-exon 1 hybrids might be digested by S1 nuclease so that the expected 249 nt signal would be reduced to about 100 nt. There is evidence in Figure 32 that the 249 bp hybrid has been digested at this A+T-rich region. A signal of about 100 nucleotides is seen. This signal from the unspliced exon 1 is not as strong as the signal from the exon 0-exon 1 form, though, particularly when RNA from liver or muscle was used in the hybridization.

It can be concluded that the two transcription and splicing pathways described by Figure 30 each generate detectable amounts of HMG-14a mRNA. The pathway represented by pLM3a (containing the 0 exon) shows the strongest S1 signal in the tissue RNAs that were examined, but the pLM2a

embryo cDNA library. However, the short 0 exon might have been lost in pLM3a-type cDNAs in this library. Such clones cannot be distinguished from truncated pLM2a-type clones. (HMG-14a cDNA clones are very rare in our liver cDNA library. Despite extensive screening only 1 independent clone of each type has been found to date.)

To show that pLM3a is a truncated cDNA clone that contains only part of exon 0, the technique of primer extension was used. In this technique, a small radioactively labelled DNA primer is hybridized to complementary sequences in a preparation of RNA. Reverse transcriptase is used to elongate the DNA, using the RNA as a template. The length of the elongated product indicates the distance from the annealing site to the 5' end of the RNA.

A 20 nucleotide primer (JD-21, see Figure 23) was designed which is complementary to exon 0 and exon 1 around the exon 0-exon 1 splice site. This primer will only anneal to mRNA containing the exon 0-exon 1 splice, and extension of this primer by reverse transcriptase will give a product indicative of the true size of exon 0. Figure 33 shows the results of primer extension with JD-21 and RNA from several chicken tissues (and yeast RNA as a negative control). A strong signal from an extended product of 139 nucleotides is seen in all 4 experimental lanes but not the control lane. Allowing for that part of the primer (and extended product)

which is complementary to exon 1, it can be deduced that exon 0 is about 126 bp, though only 12 bp of exon 0 are found in pLM3a. All of exon 0, as predicted by primer extension, is shown in Figure 27.

The HMG-14a promoter

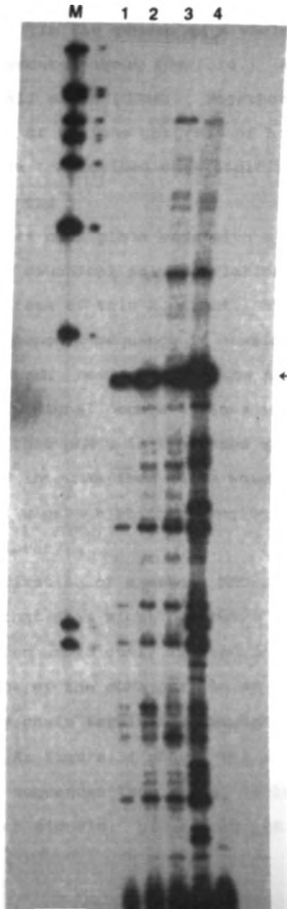
The sequences upstream of exon 1 and exon 0 are expected to contain promoter elements which facilitate transcription of the HMG-14a gene. As Figure 27 shows, the HMG-14a promoter region does not contain either the TATAA or CCAAT elements. This differs from the HMG-17 promoter, which contains two TATAA elements and a CCAAT element in a typical arrangement. The HMG-14a promoter does have some features that are found in the promoters of other genes. The SP1 binding site is found 12 times in this region. The SP1 protein and its importance for high levels of transcription have been characterized in mammalian systems. The conservation of other promoter elements in eukaryotes suggest that some or all of the SP1 binding sites could function in the HMG-14a gene. Five SP1 binding sites are clustered 34 to 62 bp upstream of exon 1; the others are scattered throughout the transcribed and untranscribed regions of the promoter.

The sequence upstream of exon 1 (Figure 27) is quite G+C-rich (76%). In this G+C-rich region the CG dinucleotide content, 115, is almost as high as the GC dinucleotide

Figure 33. Primer extension analysis of HMG-14a mRNA. The primer JD-21 (Figure 25) was used to map the initiation site of exon 0.

Experimental lanes: 1 = liver RNA
2 = brain RNA
3 = cardiac muscle RNA
4 = skeletal muscle RNA
5 = yeast RNA.

The strong 139 nucleotide signal is marked.


M
1
2
3
4
← 139

content, 127. (In the genome as a whole, CG dinucleotides are underrepresented about fourfold.) Also, this region contains 7 HpaII sites (CCGG). Together, these features are characteristic of regions upstream of housekeeping genes (19), which are transcribed constituitively in most cells.

The HMG-14a 3' end

The HMG-14a cDNA pLM3a ends with a stretch of 20 A residues. The canonical polyadenylation signal AATAAA is not found upstream of this A_{20} tract. When the corresponding genomic sequence is examined (Figure 34) the A_{20} tract is found. Downstream of the A_{20} tract the polyadenylation signal occurs 4 times in a 21 bp region. It seemed likely that pLM3a is truncated at the 3' end. This is often found in cDNAs from genes which have an A+T-rich 3' end, due to S1 cleavage at these regions during preparation of the cDNA library.

Characterization of a second HMG-14a cDNA, pLM2a, demonstrates that this might be true. pLM2a contains HMG-14a 3' sequences anomolously ligated to vector sequences. The 3' sequence of the cDNA portion of pLM2a was determined by the dideoxy chain termination method using a _gt11 primer (Figure 28). As Figure 34 shows, the 3' end of pLM2a contains more sequences than pLM3a, including the polyadenylation signals. pLM2a does not contain sequences

Figure 34. **The 3' end of the HMG-14a gene. The genomic sequence is shown. Sequences found in a cDNA clone are capitalized, the 3' ends of the 2 cDNAs analyzed are indicated with arrows.**

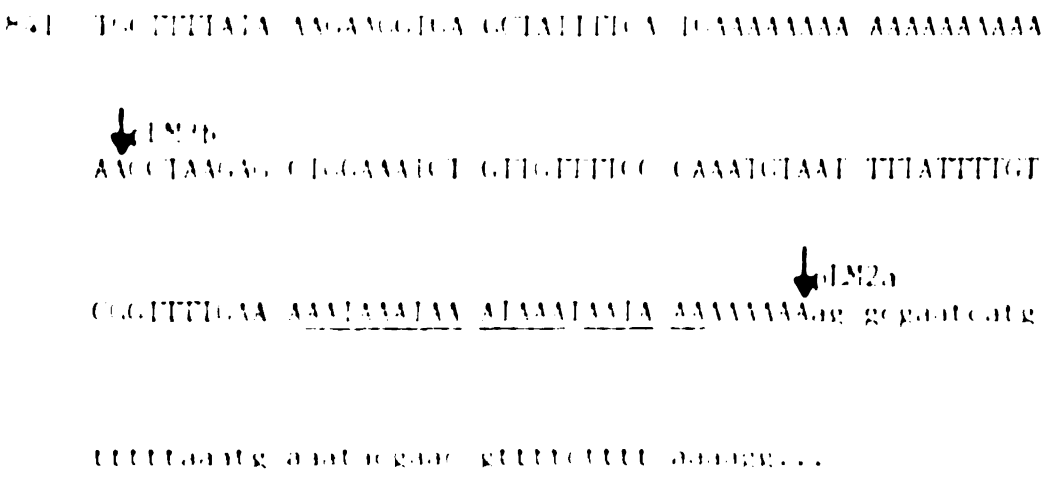


Figure 34

representative of a posttranscriptionally added polyadenylate tail, so the site of 3' processing of HMG-14a mRNA has not been unequivocally determined.

Recent characterization of 6 more HMG-14a cDNAs showed that 3 clones had 3' ends shorter than pLM3a, presumably due to truncation during cDNA cloning. Two clones ended at the same site as pLM2a, but both had longer runs of poly (dA) than are encoded at this position in the genome (A_{13} and A_{21}). This suggests that these clones, and pLM2a, end at the major polyadenylation site for HMG14a. However, one clone extended 110 bp 3' to the pLM2a end. This clone ended shortly downstream of an AATAAA signal sequence, but did not contain a run of nongenomic dAs. Two explanations for these results seem reasonable:

1. HMG-14a mRNAs terminate at more than one polyadenylation site *in vivo*; or
2. cDNA clones which end at the pLM2a site but have longer poly (dA) tracts than encoded in the genomic result from oligo dT priming at this A-rich site during cDNA synthesis. They have long dA stretches because they were primed by long oligo dT molecules and/or because of slippage during replication. This would suggest that the real polyadenylation site for HMG-14a is at least 110 bp 3' to the pLM2a end.

We favor the first of these two explanations because similar suggestions of multiple polyadenylation site usage have been found in analyzing a set of HMG-14b cDNA clones (Dodgson et al., unpublished results). It is difficult, if not impossible, to totally rule out the latter explanation, given the A+T-rich, S1-sensitive regions in the 3' untranslated region of this gene.

The HMG-14b gene

Srikantha, Landsman, and Bustin (3) have determined the structure of the HMG-14b gene. Like the HMG-14a gene, the HMG-14b gene is larger (7.8 kb) than the HMG-17 gene. Its promoter, like that of HMG-14a, contains no TATAA elements, but is G+C-rich with a relatively high CG dinucleotide content, and contains several SP1 binding sites and HpaII sites. Also like HMG-14a, it contains multiple polyadenylation signals.

An interesting difference between the two HMG-14 genes is seen when the exon-intron structure of the two genes are compared. In HMG-14b the prototypical exon 2 and exon 3 appear to be fused. This region has been highly conserved in the HMG proteins. A valine is inserted in the protein sequence, relative to the consensus sequences of human HMG-14, human HMG-17, chicken HMG-14a, and chicken HMG-17. The valine codon, GTA, may be a remnant of an intron which has been lost, since it is similar to the GT...AG sequence found at the ends of introns.

The possibility of multiple 5' splicing patterns of HMG-14b
has not been examined.

Comparison of human and chicken HMG genes

While characterization of the chicken HMG-17 and HMG-14a genes was being completed, reports of the structures of the human HMG genes were published (20,21). Comparison of the human and chicken genes shows that some features of these genes have been conserved since the division of the human and chicken gene families.

1. The HMG transcripts all have G+C-rich leaders and long A+T-rich untranslated regions.
2. Though the human and chicken HMG-17 and HMG-14 proteins are similar, the HMG-14 genes are much larger (8-10 kb) than the HMG-17 genes (3-4 kb).
3. The exon-intron structure, which is similar in the chicken HMG-17 and HMG-14 genes, is also similar in both human genes. Exon 1, exon 3, and exon 4 are particularly well conserved in the chicken genes; the human exons code for essentially identical amino acid sequences. Exon 3 and exon 4 encode the DNA binding protion of the HMG-14,17 proteins (22,23). The possibility of multiple splicing patterns in the human HMG-14 gene or chicken HMG-14b has not been investigated.

4. The promoters of both HMG-17 genes are G+C-rich and contain 2 TATAA elements, a CCAAT element, and several SP1 binding sites.
5. The promoters of both HMG-14 genes are G+C-rich. HMG-14 promoters do not contain the TATAA element but contain multiple SP1 binding sites.

We have substantially accomplished our goals in this study of HMG gene structure. This is satisfying, especially because our work, complementing the work of others, has raised new interesting questions about the proteins and their genes. We hope that the discoveries we have made will allow new investigative approaches to the study of chromatin structure and function.

References

1. Dodgson, J.B., M. Yamato, and J.D. Engel. (1987). Nuc. Acids Res. 15:6294.

2. Srikantha, T., D. Landsman, and M. Bustin. (1988). J. Biol. Chem. 263:13500.

3. Srikantha, T., D. Landsman, and M. Bustin. (1990). J. Mol. Biol., in press.

4. Lewin, B. (1980). "Gene Expression, Vol. 2", p. 816. Wiley, New York.

5. Proudfoot, N.J, and G.G. Brownlee. (1974). Nature (London) 252:359.

6. Aho, S., U. Tate, and H. Boedtker. (1983). Nuc. Acids Res. 11:5443.

7. Tosi, M., R.A. Young, O. Hagenbuchle, and U. Schibler. (1981). Nuc. Acids Res. 9:2313.

8. Goldberg, M. (1979). Thesis. Stanford University.

9. Hu, S.L., and J.L. Manley. (1981). Proc. Nat. Acad. Sci. USA (USA) 78:820.

10. Fletcher, C., N. Heintz, and R.G. Roeder. (1987). Cell 51:773.

11. McKnight, S.L., and R. Kingsbury. (1982). Science 217:316.

12. Hentschel, C.C., and M.L. Birnstiel. (1981). Cell 25:301.

13. Kadonaga, J.T., K.A. Jones, and R. Tijan. (1986). Trends in Biochem. Sci. 11:20.

14. Maniatis, T., S. Goodbourn, and J.A. Fischer. (1987). Science 236:1237.

15. Dynan, W.S., and R. Tijan. (1983). Cell 32:669.

16. Brush, D., J.D. Dodgson, O. Choi, P. Stevens, and J. Engel. (1985). Mol. Cell. Biol. 5:1307.

17. Landsman, D., T. Srikantha, and M. Bustin. (1988). J. Biol. Chem. 263:3917.

18. Young, R.A., and R.W. Davis. (1983). Proc. Nat. Acad. Sci. USA 80:1194.

19. Bird, A.P. (1986). Nature (London) 321:209.

20. Landsman, D., O.W. McBride, and M. Bustin. (1989). Nuc. Acids Res. 17:2301.

21. Landsman, D., O.W. McBride, N. Soares, M. Crippa, T. Srikantha, and M. Bustin. (1989). J. Biol. Chem. 264:3421.

22. Javaherian, K., and S. Amini. (1978). Biochem. Biophys. Res. Com. 85:1385.

23. Abercrombie, B.D., G.G. Kneale, C. Crane-Robinson, E.M. Bradbury, G.H. Goodwin, J.M. Walker, and E.W. Johns. (1978). Eur. J. Biochem. 84:173.

APPENDIX

Gene, 63 (1988) 287–295
Elsevier

287

GEN 02323

Chicken chromosomal protein HMG-14 and HMG-17 cDNA clones: isolation, characterization and sequence comparison

(Recombinant DNA; oligodeoxynucleotide screening; gene copy number; mRNA levels)

Jerry B. Dodgson[a,b], David L. Browne[b] and Andrew J. Black[a]

Departments of [a] Microbiology and Public Health and [b] Biochemistry, Michigan State University, East Lansing, MI 48824 (U.S.A.) Tel. (517)355-6464

Received 6 October 1987
Accepted 3 December 1987
Received by publisher 31 December 1987

SUMMARY

A cDNA clone coding for the chicken high-mobility group 14 (HMG-14) mRNA has been isolated from a chicken-liver cDNA library by screening with two synthetic oligodeoxynucleotide pools whose sequences were derived from the partial amino acid sequence of the HMG-14 protein. A chicken HMG-17 cDNA clone was also isolated in a similar fashion. Comparison of the two chicken HMG cDNA clones to the corresponding human cDNA sequences shows that chicken and human HMG-14 mRNAs and polypeptides are considerably less similar than are the corresponding HMG-17 sequences. In fact, the chicken HMG-14 is almost as similar to the chicken HMG-17 in amino acid sequence as it is to mammalian HMG-14 polypeptides. HMG-14 and HMG-17 mRNAs seem to contain a conserved sequence element in their 3'-untranslated regions whose function is at present unknown. The chicken HMG-14 and HMG-17 genes, in contrast to their mammalian counterparts, appear to exist as single-copy sequences in the chicken genome, although there appear to exist one or more additional sequences which partially hybridize to HMG-14 cDNA. Chicken HMG-14 mRNA, about 950 nucleotides in length, was detected in chicken liver RNA but was below our detection limits in reticulocyte RNA.

INTRODUCTION

The HMG proteins are small non-histone chromosomal proteins, rich in both basic and acidic amino acids, which appear to be widely distributed, if not ubiquitous, in eukaryotes (Johns, 1982). In general, the function(s) of HMG proteins remains unknown, but Weisbrod and Weintraub (1979) presented evidence indicating that the chicken HMG-14 and HMG-17 proteins appear to be necessary components of actively transcribing chromatin. Landsman, Bustin and colleagues have recently

Correspondence to: Dr. J.B. Dodgson, Department of Microbiology and Public Health, Giltner Hall, Michigan State University, East Lansing, MI 48824 (U.S.A.) Tel. (517)353-5024.

Abbreviations: aa, amino acid(s); bp, base pair(s); cDNA, DNA complementary to RNA; HMG, high mobility group; I, inosine; kb, 1000 bp; nt, nucleotide(s); pfu, plaque-forming units; SDS, sodium dodecyl sulfate; SSC, 0.15 M NaCl, 0.015 M Na_3·citrate, pH 7.5; TMAC, tetramethylammonium chloride; ts, temperature sensitive.

obtained cDNA clones coding for the human HMG-14 (Landsman et al., 1986b) and HMG-17 (Landsman et al., 1986a) mRNAs and provided evidence for the existence of multigene families in both cases (Landsman and Bustin, 1986).

We have isolated chicken HMG-14 and HMG-17 cDNA clones from a chicken liver cDNA library using hybridization with synthetic oligodeoxynucleotide pools whose sequences were determined from the known complete amino acid sequence of chicken HMG-17 (Walker et al., 1980) and the partial sequence of chicken HMG-14 (Walker, 1982). Because one stretch of sequence is identical in both HMG proteins, one of the oligodeoxynucleotide preparations hybridized to both cDNAs. HMG-14-specific and HMG-17-specific oligodeoxynucleotide probes were also prepared. At degenerate positions in the sequence, either all possible bases were used or deoxyinosine was used since it has been shown, in at least some cases, to provide a useful alternative to making a large number of different oligomers (Ohtsuka et al., 1985).

The complete sequences of both HMG clones were determined. The chicken HMG-17 sequence is nearly identical to one described very recently by Landsman and Bustin (1987). The HMG-14 sequence is considerably different from that of human HMG-14. Furthermore, the HMG-14 and HMG-17 genes appear to be single copy in the chicken genome in contrast to the situation in man and mammals.

MATERIALS AND METHODS

(a) Oligodeoxynucleotide screening

The following oligodeoxynucleotides were purchased from the Michigan State University Macromolecular Structure Facility:

d(GCYTTYTTIGGYTTIGGYTCIGGYTTIGG), complementary to both HMG-14 and HMG-17 sequences;

d(TCYTTYTTRTCRTCIGCIGCYTTYTCYTT), complementary to HMG-14; and

d(TGIGGYTCRTCYTTNACYTTIGCYTTRTC), complementary to HMG-17.

Y indicates that an equimolar combination of pyrimidines was used in the synthesis; R indicates the use of equimolar purines, and N the use of all 4 nt. Deoxyinosine was chosen for incorporation at four-fold ambiguous positions where a review of codon usage in other sequenced chicken genes showed that it was unlikely that an I would be directly opposite a G in the cDNA.

Oligodeoxynucleotides were labeled with polynucleotide kinase and [γ-^{32}P]ATP as described (Maniatis et al., 1982). The chicken liver λgt11 cDNA library used has been described previously (Dodgson et al., 1987) and was a gift of M. Yamamoto and J.D. Engel (Northwestern University). HMG cDNA clones were isolated by the technique of Jacobs et al. (1985) in which 3 M TMAC is used to make hybridization temperature strictly dependent on oligodeoxynucleotide length. Phage were plated at 10–20000 pfu/150 mm plate, filter (Dupont Corp., NEF-978A) replicates were made, and the plaques were amplified by the method of Woo (1979). The filters were processed as described (Woo, 1979) and extensively prewashed in 3 M TMAC; 0.05 M Na phosphate, pH 6.8; 1 mM EDTA; 0.5% SDS; 0.1% polyvinylpyrrolidone; 0.1% Ficoll; 0.1% bovine serum albumin and 100 μg/ml of denatured salmon sperm DNA. After overnight prehybridization in the same solution at 53–55°C, labeled oligodeoxynucleotides (2–10 x 10^6 cpm/pmol) were added to about 0.2 pmol oligomer/ml. Hybridization continued at 53–55°C for 48 h. Filters were washed several times in 2 x SSC at 25°C, followed by washes in 3 M TMAC; 0.05 M sodium phosphate, pH 6.8; 0.1% SDS; 1 mM EDTA; one at 25°C and two at 45–50°C. Positive plaques were picked and rescreened at least one more time. For the HMG-14 screen four positive clones were identified from about 10^5 initial phage, but only one of these had an insert which could be directly excised with *Eco*RI, so only this clone was studied further. For the HMG-17 screen one clone was obtained from about 10^4 initial phage.

(b) Miscellaneous techniques

cDNA inserts were cloned into pT3/T7-mp18 (Bethesda Research Labs, Gaithersberg, MD) and restriction maps of the resulting subclones prepared by standard techniques (Maniatis et al., 1982). Nucleotide sequence analysis was done by the

Maxam and Gilbert (1980) method as described (Grandy and Dodgson, 1987). Genomic DNA and RNA blotting experiments and nick translations were performed as described previously (Grandy and Dodgson, 1987, Maniatis et al., 1982). Nucleotide comparison analysis was done using the GENEPRO program (Riverside Scientific Enterprises, Seattle, WA).

RESULTS AND DISCUSSION

(a) HMG-14 and HMG-17 nucleotide sequence

The sequence of the chicken HMG-14 cDNA insert is shown in Fig. 1, along with the derived HMG-14 polypeptide sequence. The insert is 900 bp in length including two linkers and 20 bp of cDNA corresponding to the poly(A) region of the mRNA. It contains 104 bp of 5'-untranslated sequence and a relatively long 3'-untranslated region of 445 bp. The chicken HMG-14 polypeptide sequence predicts a protein of 104 aa residues. This is less than the 121 aa residues predicted by peptide analysis (Walker, 1982) but more closely resembles the 98-aa human HMG-14 (Landsman et al., 1986b) and 100-aa calf HMG-14 (Walker et al., 1979). The protein encoded by the chick HMG-14 cDNA has two differences from the reported partial amino acid sequence of Walker (1982): it contains lysine rather than proline at aa position 41 and asparagine rather than aspartic acid at 47. However, the overall amino acid composition of the predicted chick HMG-14 protein agrees very closely with the experimental data (Walker and Johns, 1980). The sequence of the chicken HMG-17 cDNA insert was also obtained (Fig. 2). It is virtually identical to the sequence published recently by Landsman and Bustin (1987) except that it contains 3 bp more of the 5'-untranslated region and appears to contain only 7 T nt in a stretch from 549–555 (Fig. 2) instead of 8 T's as their sequence contains.

(b) Sequence comparison to human HMG proteins and mRNAs

A dot-matrix (Maizel and Lenk, 1981) comparison of the chicken HMG-14 amino acid se-

```
  GAATTCCGTC CCCTTCCTCA GGACGCTCGA AAACAGTTTC TCGCGGGTTC CCTTCCTATT    60
  TTTTACACCT CTCCCGATCT CTCTATTTGC AGTCAACTAT TAAGGTGCAA CTATGCCCAA   120
                                                              ProLy
  AAGAAAGCCT CCAGCTGAAG GCGAGGCGAA GGAGGAGCCA AAGAGAAGGT CGGCCAGACT   180
2 sArgLysAla ProAlaGluG lyGluAlaLy sGluGluPro LysArgArgS erAlaArgLe
  ATCTGCTAAA CCTGCTCCGC CTAAACCGGA GCCAAAGCCC AAAAAGGCAG CACCTAAGAA   240
22 uSerAlaLys ProAlaProP roLysProGl uProLysPro LysLysAlaA laProLysLy
  AGAAAAGGCA GCAAACGATA AAAAGGAAGA CAAAAAGGCA GCAACAAAAG GCAAGAAAGG   300
42 sGluLysAla AlaAsnAspL ysLysGluAs pLysLysAla AlaThrLysG lyLysLysGl
  AGCCAAAGGC AAAGACGAAA CTAAACAAGA GGATGCAAAA GAAGAAAACC ACTCTGAAAA   360
62 yAlaLysGly LysAspGluT hrLysGlnGl uAspAlaLys GluGluAsnH isSerGluAs
  TGGAGATACC AAAACTAATG AGGCACCAGC TGCTGAAGCA TCTGATGATA AGGAAGCCAA   420
82 nGlyAspThr LysThrAsnG luAlaProAl aAlaGluAla SerAspAspL ysGluAlaLy
  GTCCGAGTAA TGTTAACCCT GCCCTATATC TCCATCATTT GGTATCCGTA CCTCCATGCT   480
102 sSerGlu***
  GTATTGTTAA CAGAGAGGAA TATTTTTATC AACTATTTTA TAAATGCAGG TTTTTTTAGC   540
  ATGAATTTAA TTATGGAACA TCTTCATCTC GGTTACTTGG GAATTAAATC CCTAACAAAC   600
  AAAACAAAAC AAAACAAAAA AAAATCATTG TTTTAAATTT GTGATTGTAA TAGTTTGTAT   660
  GGTACATGGA AAGAATAAGT GGTGGTAGCT TTTGACTTCT GTCAGTGTGT CCCTTTTTGT   720
  GTAAGTCATC CTTACAGACT TCAGATTTTA ATTTTACCCT TGTATGTGTT GTATGGTTTC   780
  TTAAAGTGGG GAGGTCTCAA AACAGATAAC TGTGTTAAAC ATTCCAGTGG TTGTGTGGGT   840
  TGCTTTTATA AAGAAGGTGA GCTATTTTCA TGAAAAAAAA AAAAAAAAAA AACGGAATTC   900
```

Fig. 1. Nucleotide sequence of chicken HMG-14 cDNA. The sequence of the chicken HMG-14 cDNA (in plasmid pLM3B) is shown with the HMG-14 protein sequence placed under the appropriate coding region. The seven bases at each end of the sequence arise from synthetic *Eco*RI linkers. Numbers on the right specify the nucleotide position and on the left the amino acid position in the sequence.

```
    GAATTCCGCA GCCAGCGCAG CGAGCCGCCC GCCAGCCCCG CCGCGCCGCC CCGCTCTCCC    60
    CCTCGGCCCT CCCCCGCTTC TCGCCGCCAC CGAGCGAGCC CGGCTGCCCG CCCCCGCCCG   120
    CCCCCTCCGC TCGCTCTCTC CCTCCTCGCA CAACACACGC ACGCGCCGCC CGGAGCTATG   180
    CCGAAGAGAA AGGCTGAAGG AGATACCAAG GGCGATAAGG CCAAAGTTAA GGATGAGCCA   240
  1 ProLysArgL ysAlaGluGl yAspThrLys GlyAspLysA laLysValLy sAspGluPro
    CAACGGAGAT CGGCAAGGTT ATCTGCTAAA CCTGCCCCTC CGAAGCCAGA GCCTAAACCT   300
 21 GlnArgArgS erAlaArgLe uSerAlaLys ProAlaProP roLysProGl uProLysPro
    AAAAAGCCAG CTCCAAAGAA GAGTGAGAAG GTGCCCAAGG GAAAGAAGGG GAAAGCTGAT   360
 41 LysLysAlaA laProLysLy sSerGluLys ValProLysG lyLysLysGl yLysAlaAsp
    GCTGGCAAGG ACGCAAACAA CCCTGCAGAA AATGGAGATG CCAAAACAGA CCAGCCACAG   420
 61 AlaGlyLysG luGlyAsnAs nProAlaGlu AsnGlyAspA laLysThrAs pGlnAlaGln
    AAAGCCGAAG GTGCTGGTGA TGCCAAGTAA AATGTGTGAA TTTTTGATAA CTGTGTACTT   480
 81 LysAlaGluG lyAlaGlyAs pAlaLys***
    CTGGTGACTG TACAGTTTGA AATACTATTT TTTATCAAGT TTTATAACAA TGCAGAATTT   540
    TGTTTTACTT TTTTTAAGCT ATGTTGTTAG CACACAGACC GCTTTGTTGT TGTGTTTTGA   600
    GGGCGGCGCA GTGCGACAAA CGTCACTTAA TCTGTTTCTT CGAACCTAAA TTTTAAAAGT   660
    TTACCCCTTC CCAGTTTTTT AGAAGGACTC TTCCTAAATG GAGCAGCAAG GGATTCCTTC   720
    GTGCTGCACA CCTCTTCCGT TTTGTGGACC GCATCAGAGT GAACGGAAGC TCCCGAGATG   780
    CCTGTTGCCA ACTTCAGAAC TGCAGTTTGC AGTGCCCTCT GCGTTTCCTT TCATGCCCTC   840
    CCTTTTTGCC TAGAGCCTAT CACTCCGAAA TACAGCAGAC ATGGCATGTT GGGACTCACC   900
    ACTCTAAATG CATTGTCAGG TGATCTGGAC TTCTGGTGTC TAATTTGCGA TATAATAGCT   960
    CTAAAAGGAG CTGCATTTCC TCTTTCATAT TGTAGATCTA CAGATTAAGG AATCTGCAGT  1020
    TTTTAATTTT TCCTCGCAAA GCTAGGGTAG ATTTGTGAAG AGTTGTTAAA CAACATGCTA  1080
    AATGTGAAAG TGTCCGCCCT CACTCTAAAC ATTTCCCTCT ACAAGTATAC AAAAATGAAG  1140
    ATTTGTCGGT TTTATAGCAA CCTTTATGTT TGGGTAGTCC ATGAAGGGAG GGGAGTTTGA  1200
    CAGTTCTTGT AAAATGTTGC AGATTGTAGC CCATGTCCTG CCTAAATTAC CATGATTGTT  1260
    TATGAAAAGT ACCTTTAATA AAGCTGGATA CGGTTTGGCT TGGAAAAAAA AAAAAAAAAA  1320
    AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AACGCAATTC                        1360
```

Fig 2 Nucleotide sequence of the chicken HMG-17 cDNA The sequence of the chicken HMG-17 cDNA insert (in plasmid pLG1A) is shown in the same manner as for Fig 1

quence to the human HMG-14 sequence is shown in Fig 3A. The strong diagonal line demonstrates significant sequence similarity throughout the molecule, and especially from aa 10 to aa 25. However, the overall similarity in sequence between the two is only 51%. For comparison, chicken HMG-17 shows 94% similarity to the corresponding human protein. It is also of interest to compare the chicken HMG-14 and HMG-17 sequences to each other. As shown in Fig 3B and as previously pointed out by Walker (1982), the chicken HMG-14 sequence is more similar to HMG-17 (either chicken or calf) in the region from aa 14 to 37 (19 to 42 in HMG-17) than it is to the mammalian HMG-14s. Overall, the two sequences are 44% identical. As for the corresponding human HMGs, the region of highest similarity between the chicken HMG-14 and HMG-17 proteins corresponds to their respective DNA-binding domains (Landsman and Bustin, 1986).

Fig 4 compares the human and chicken HMG-14 and HMG-17 mRNA (or cDNA) sequences to each other. Fig 4A compares chicken HMG-17 cDNA to the corresponding human cDNA. Clearly, sequence similarity extends throughout the length of the mRNAs. Not surprisingly, the similarity is especially strong in the coding portion of the molecule, but it is also very strong in the 3' region of the mRNAs, perhaps reflecting sequences required for proper mRNA polyadenylation and or mRNA structure and stability. Fig 4B shows the analogous comparison of chicken versus human HMG-14 mRNAs. Again, there is much less interspecies similarity in the HMG-14 family. First of all, the chicken HMG-14 sequence is 340 nt shorter, due to a shorter 3'-untranslated region. What similarity does exist is primarily in the coding region. However, there is also very strong similarity between the two cDNAs at nt positions 500 to 550 (in both sequences). This begins

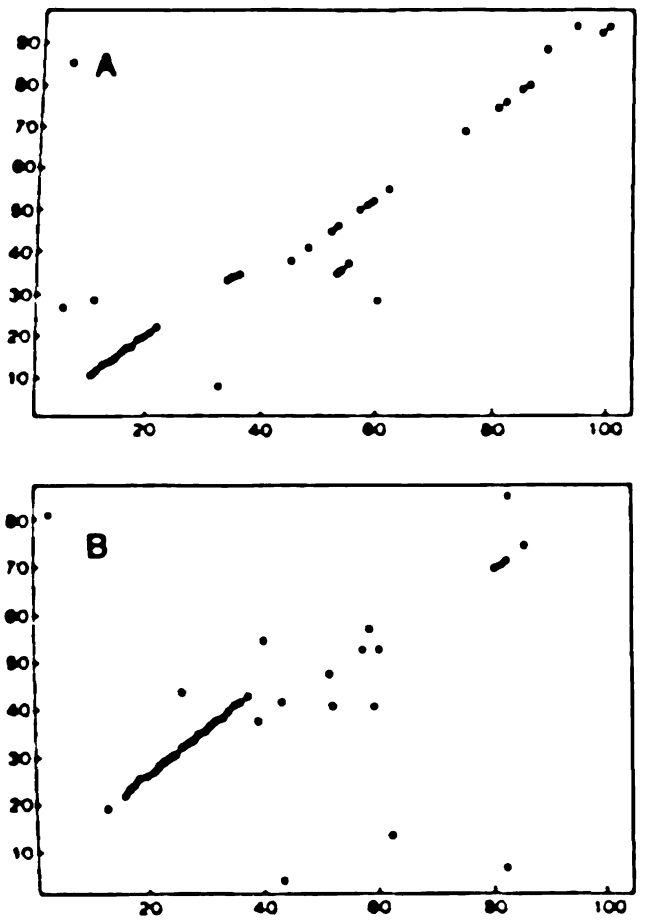
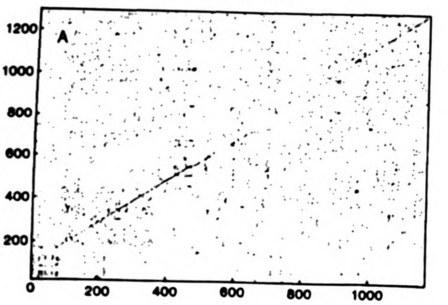


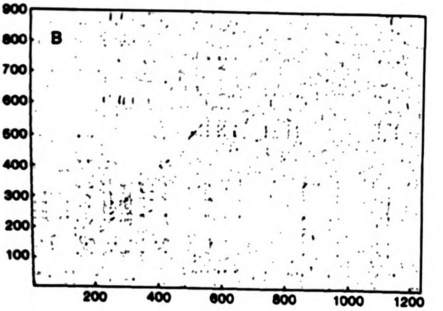
Fig 3 Dot matrix comparison of HMG-14 amino acid sequence The chicken HMG-14 amino acid sequence (X axis) is compared to that of human HMG-14 (Y axis, panel A) and chicken HMG-17 (Y axis, panel B) The human HMG-14 sequence is from Landsman et al (1986b) The window of comparison was 10 aa residues with 50% identity required for a positive result to be recorded

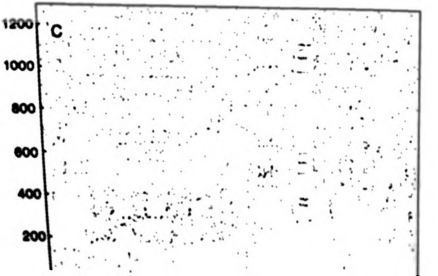
about 50 nt 3′ to the stop codon in the chicken HMG-14 sequence (40 nt 3′ in the human) The function of this region is unknown, but the high level of sequence similarity in this area strongly suggests an important role in the HMG-14 mRNA structure This sequence may be important in both the HMG-14 and HMG-17 mRNAs, since a comparison of the chick HMG-17 cDNA to that of the chick HMG-14 (Fig 4C) also shows similarity between the two regions at nt positions 500–550 The level of sequence similarity between these two cDNAs in this region is not as great as exists between the chicken and human HMG-14 cDNAs, but it is of the same magnitude as the similarity between the chick HMG-14 and HMG-17 coding regions which, as described above, leads to 44% identity in their amino acid sequences Fig 4C also identifies a region very rich in A residues (nt 594–624 in Fig 1) in the chick HMG-14 3′-untranslated region This shows up in Fig 4C as a series of horizontal lines which mark the similarity of this tract to several smaller A-rich regions in HMG-17. (The A blocks corresponding to the poly(A) tails have been deleted in Fig 4.) Both chicken cDNAs share the characteristics noted for the human HMG-14 and HMG-17 cDNAs (Landsman et al., 1986a,b) of being G + C-rich in their 5′-untranslated regions and A + T-rich in their 3′-untranslated regions The functions, if any, of these nucleotide biases and the long A tract are unknown

(c) Gene copy number

Fig 5 shows blots of chicken genomic DNA cut with either *Eco*RI or *Bam*HI restriction enzymes and hybridized with the HMG-14 or HMG-17 cDNA inserts In each case, only one strongly hybridizing band is seen, suggesting that both cDNA sequences are single copy in the chicken genome. This has been confirmed by isolation of cloned genomic DNAs (D L B , unpublished results) whose restriction maps demonstrate that the strongly hybridizing bands do not result from HMG genes duplicated in tandem or several HMG genes closely linked on a single restriction fragment For both HMG genes, approximately one positive clone was isolated per 50000 λ recombinants (15–20 kb inserts) screened, in agreement with each gene being single copy in the chicken genome. However, in the case of HMG-14 there is one other band which hybridizes at 20–30%, the strength of the major band (at 7 8 kb in lane 2 and 6 9 kb in lane 4 of Fig 5) and one or two other still weaker bands. Preliminary results (D L B) from our genomic clones suggest that these minor bands are not due to small portions of the HMG-14 gene existing as separate exons on different restriction fragments. The weaker bands are likely to result from partial sequence similarity of the HMG-14 probe to other sequences in the chicken genome, perhaps to other HMG genes. However, despite the significant similarity of chicken HMG-14 and HMG-17 coding sequences, there is no observable cross-hybridization of the two probes under the conditions of this experiment (Fig. 5, lanes 1 and 2).


A
1200
1000
800
600
400
200
0
200
400
600
800
1000


B
900
800
700
600
500
400
300
200
100
200
400
600
800
1000
1200


C
1200
1000
800
600
400
200

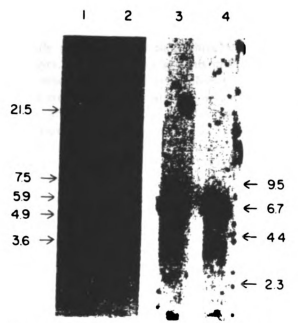


Fig. 5. Chromosomal blots of chicken genomic DNA hybridized to HMG cDNAs. Each lane contained 11 μg of chicken genomic DNA digested with *Eco*RI (lanes 1, 2) or *Bam*HI (lanes 3, 4). Blots were hybridized in 50% formamide hybridization solution at 42°C as described (Grandy and Dodgson, 1987) to nick-translated (10^8 cpm/μg) [^{32}P]cDNA inserts from the HMG-17 (lanes 1, 3) or the HMG-14 (lanes 2, 4) clone. Blots were washed at 65°C in 0.1 M NaCl–0.01 M Tris·HCl, pH 7.5–1 mM EDTA. Arrows denote the positions and sizes in kb of internal *Eco*RI-digested λ DNA markers (lanes 1, 2) or external *Hin*dIII-digested λ DNA markers (lanes 3, 4).

Since it appears that the genes for HMG-14 and HMG-17 are single copy in the chicken genome, it is possible that most of the multigene family members observed for these two cDNAs in man and mammals (Landsman and Bustin, 1986) are pseudogenes. For unknown reasons pseudogenes seem to be rather rare in the chicken genome. For example, we have yet to identify a single pseudogene in either of the two chicken globin gene clusters (three and four genes each; Dodgson et al., 1981; Dolan et al., 1981) or the two replication variant histone gene clusters (19 and 21 genes each; Grandy and Dodgson, 1987).

(d) HMG-14 mRNA levels

Preliminary measurements have been made of chicken HMG-14 mRNA levels by RNA blotting as shown in Fig. 6. As expected from our cloning results (four positives from about 100000 phage, see MATERIALS AND METHODS, section **a**), there is a low but clearly measurable level of HMG-14 mRNA in chicken liver total RNA (Fig. 6, lane 3). A single band was observed of a size (approx. 950 nt) similar to that of the cloned cDNA insert. However, we were unable to detect HMG-14 mRNA in either reticulocyte RNA from anemic birds or in RNA from HD3 cells, an erythroid precursor cell line (transformed with *ts* avian erythroblastosis virus; Beug et al., 1982) grown in culture. HMG-14 and HMG-17 are

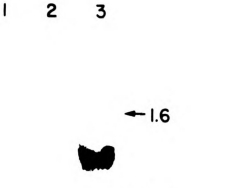


Fig. 6. Chicken HMG-14 mRNA levels. Total RNA was prepared as described (Yoshihara et al., 1987) from: lane 1, HD3 chicken cells (Beug et al., 1982); lane 2, anemic hen reticulocytes; and lane 3, adult chicken liver. RNA samples (100 μg/lane) were run on a 2.2-M formaldehyde–1.2% agarose gel and the gel was blotted as described (Maniatis et al., 1982). Hybridization with nick-translated HMG-14 cDNA and washing were as described in the legend to Fig. 5. The arrows at 600 and 1600 nt designate the positions of an internal RNA (α-globin mRNA) and an external single-stranded DNA standard, respectively. RNA and DNA standards were shown to run equivalently on this gel.

Fig. 4. Dot-matrix comparison of HMG cDNA sequences. (**A**) Comparison of human HMG-17 cDNA (X-axis) to chicken HMG-17 cDNA (Y-axis). (**B**) Comparison of human HMG-14 cDNA (X-axis) to chicken HMG-14 cDNA (Y-axis). (**C**) Comparison of chicken HMG-14 cDNA (X-axis) to chicken HMG-17 cDNA (Y-axis). Human cDNA sequences are from Landsman et al. (1986a,b). A window of 10-nt residues was used with 50% identity required for a positive result in all cases. Linker and 3′ poly(A) regions have been removed for this analysis.

clearly present in chicken erythrocyte chromatin (Mayes, 1982), but these proteins may have been synthesized early in erythroid differentiation and/or were translated from relatively low mRNA levels, thus accounting for our inability to detect the message in reticulocyte mRNA. The absence of detectable message in HD3 cells is surprising in view of the results of Bustin et al. (1987), which showed much higher HMG-17 mRNA levels in cultured cells than in liver, but these authors also found considerably higher levels of HMG-17 mRNA than HMG-14 mRNA in HeLa cells. More sensitive measurements will be required to delineate the overall regulation of chicken HMG-14 mRNA levels.

(e) Conclusions

Nucleotide sequence analysis of chicken HMG-14 and HMG-17 cDNAs demonstrates considerable sequence similarity between avian and mammalian HMG-17 sequences but much less similarity between the analogous HMG-14 sequences. However, comparison of several HMG-14 and HMG-17 cDNA sequences suggests a potential conserved regulatory region in the 3'-untranslated portion of these mRNAs. In contrast to the mammalian HMG-14 and HMG-17 gene families, these sequences appear to exist in one to two copies per haploid chicken genome. Low levels of the HMG-14 mRNA, 950 nt in length, were detected in total chicken liver RNA but not in RNA isolated from anemic chicken reticulocytes or from a chicken erythroblast cell line grown in culture.

ACKNOWLEDGEMENTS

We thank Drs. Ed Fritsch (Genetics Institute, Cambridge, MA) and Ron Davis (Michigan State University) for discussions regarding oligodeoxynucleotide screening techniques. This work was supported by an All-University Research Initiation Grant from Michigan State University and by an NIH Grant (GM 28837) to J.B.D.; J.B.D. is the recipient of a Research Career Development Award from NIH. This is journal article No. 12428 of the Michigan Agricultural Experiment Station.

REFERENCES

Beug, H., Palmieri, S., Freudenstein, C., Zentgraf, H. and Graf, T.: Hormone-dependent terminal differentiation in vitro of chicken erythroleukemia cells transformed by *ts* mutants of avian erythroblastosis virus. Cell 28 (1982) 907–919.

Bustin, M., Soares, N., Landsman, D., Srikantha, T. and Collins, J.M.: Cell cycle regulated synthesis of an abundant transcript for human chromosomal protein HMG-17. Nucl. Acids Res. 15 (1987) 3549–3561.

Dodgson, J.B., McKune, K.C., Rusling, D.J., Krust, A. and Engel, J.D.: Adult chicken α-globin genes, α^A and α^D: no anemic shock α-globin exists in domestic chickens. Proc. Natl. Acad. Sci. USA 78 (1981) 5998–6002.

Dodgson, J.B., Yamamoto, M. and Engel, J.D.: Chicken histone H3.3B cDNA sequence confirms unusual 3' UTR structure. Nucl. Acids Res. 15 (1987) 6294.

Dolan, M., Sugarman, B.J., Dodgson, J.B. and Engel, J.D.: Chromosomal arrangement of the chicken β-type globin genes. Cell 24 (1981) 669–677.

Grandy, D.K. and Dodgson, J.B.: Structure and organization of the chicken H2B histone gene family. Nucl. Acids Res. 15 (1987) 1063–1080.

Jacobs, K., Shoemaker, C., Rudersdorf, R., Neill, S.D., Kaufman, R.J., Mufson, A., Seehra, J., Jones, S.S., Hewick, R., Fritsch, E.F., Kawakita, M., Shimizue, T. and Miyake, T.: Isolation and characterization of genomic and cDNA clones of human erythropoietin. Nature 313 (1985) 806–810.

Johns, E.W. (Ed.): The HMG Chromosomal Proteins. Academic Press, New York, 1982.

Landsman, D. and Bustin, M.: Chromosomal proteins HMG-14 and HMG-17. J. Biol. Chem. 261 (1986) 16087–16091.

Landsman, D. and Bustin, M.: Chicken non-histone chromosomal protein HMG-17 cDNA sequence. Nucl. Acids Res. 15 (1987) 6750.

Landsman, D., Soares, N., Gonzales, F.J. and Bustin, M.: Chromosomal protein HMG-17. J. Biol. Chem. 261 (1986a) 7479–7484.

Landsman, D., Srikantha, T., Westermann, R. and Bustin, M.: Chromosomal protein HMG-14. J. Biol. Chem. 262 (1986b) 16082–16086.

Maizel, Jr., J.V. and Lenk, R.P.: Enhanced graphic matrix analysis of nucleic acid and protein sequences. Proc. Natl. Acad. Sci. USA 78 (1981) 7665–7669.

Maniatis, T., Fritsch, E.F. and Sambrook, J.: Molecular Cloning. A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1982.

Maxam, A.M. and Gilbert, W.: Sequencing end-labeled DNA with base-specific chemical cleavage. Methods Enzymol. 65 (1980) 499–560.

Mayes, E.L.V.: Species and tissue specificity. In Johns, E.W. (Ed.), The HMG Chromosomal Proteins. Academic Press, New York, 1982, pp. 9–40.

Ohtsuka, E., Matsuki, S., Ikehara, M., Takahashi, Y. and Matsubara, K.: An alternative approach to deoxyoligonucleotides as hybridization probes by insertion of deoxyinosine at ambiguous codon positions. J. Biol. Chem. 260 (1985) 2605–2608.

Walker, J.M.: Primary structures. In Johns, E.W. (Ed.), The HMG Chromosomal Proteins. Academic Press, New York, 1982, pp. 69–87.

Walker, J.M. and Johns, E.W.: The isolation, characterization and partial sequences of the chicken erythrocyte non-histone chromosomal proteins HMG14 and HMG17. Biochem. J. 185 (1980) 383–386.

Walker, J.M., Goodwin, G.H. and Johns, E.W.: The primary structure of the nucleosome-associated chromosomal protein HMG14. FEBS Lett. 100 (1979) 394–398.

Walker, J.M., Stearn, C. and Johns, E.W.: The primary structure of non-histone chromosomal protein HMG17 from chicken erythrocyte nuclei. FEBS Lett. 112 (1980) 207–210.

Woo, S.L.C.: A sensitive and rapid method for recombinant phage screening. Methods Enzymol. 68 (1979) 389–395.

Yoshihara, C.M., Lee, J.-D. and Dodgson, J.B.: The chicken carbonic anhydrase II gene: evidence for a recent shift in intron position. Nucl. Acids Res. 15 (1987) 753–770.

Communicated by J.L. Slightom.