Volume 13 Number 13 1985 Nucleic Acids Research

A comparison between mammalian and avian fast skeletal muscle alkali myosin light chain genes: regulatory implications

Philippe Daubas, Benoît Robert, Ian Garner and Margaret Buckingham

Department of Molecular Biology, Institut Pasteur, 25 rue du docteur Roux, 75724 Paris Cedex 15, France

Received 17 May 1985; Accepted 4 June 1985

ABSTRACT

A single locus in the mouse, rat and chicken encodes both alkali myosin light chains, $MLC1_F$ and $MLC3_F$. This gene has two distinct promoters and gives rise to two different primary transcripts, which are processed by alternative and different modes of splicing to form $MLC1_F$ and $MLC3_F$ mRNAs. The $MLC1_F/MLC3_F$ gene is very similar between mouse, rat and chicken, in terms of its overall structure, the length and location of the introns, and the splice site consensus sequences. Nucleotide sequences of coding regions are very conserved but 3' and 5' non coding regions of the mRNAs have diverged . In the $MLC1_F$ promoter regions, several blocks of nucleotides are highly conserved (more than 70 % homology), especially a sequence of about 70 nucleotides, located between positions -80 and -150 relative to the Cap site. Conserved blocks of homology are also found in the $MLC3_F$ promoter regions, although the common sequences are shorter. The presence of such highly conserved nucleotide sequences in the 5' flanking regions suggests that these sequences are functionally important in initiation of transcription and regulation of expression of this complex gene. Primer extension experiments indicate multiple cap sites for $MLC3_F$ mRNA.

INTRODUCTION

The two alkali myosin light chains ($MLC1_F$ and $MLC3_F$) expressed in adult fast skeletal muscle are encoded by a single gene in mammals (1,2) and birds (3). This was originally suggested by aminoacid sequence data which showed that the $MLC1_F$ and $MLC3_F$ proteins of rabbit skeletal muscle share a common COOH-terminal sequence of 141 aminoacids and have distinct NH_2-termini of 50 and 8 aminoacids for $MLC1_F$ and $MLC3_F$ respectively (4), taking into account the N-terminal α-N trimethylalanine (5) in the case of $MLC1_F$. Furthermore, in contrast to the situation for their NH_2-terminal sequences, any aminoacid substitution occuring in the COOH-terminal region of $MLC1_F$ between avian and mammalian species is found at the same position in $MLC3_F$ (6).

We have demonstrated that there is one functional genetic locus for these proteins in the mouse, although in some mouse species a second locus has been characterized which corresponds to a processed pseudogene (1). DNA

sequencing of recombinant phages containing this gene reveals its complex and unusual structure : the COOH-terminal coding (141 aminoacids) and 3' non coding regions of $MLC1_F$ and $MLC3_F$ are represented in 5 common exons, while the specific NH_2 termini are encoded by a further 4 distinct 5' exons, two for $MLC1_F$ localised on either side of a pair for $MLC3_F$ (figure 1). S1 protection and primer extension experiments suggest separate Cap sites, and promoter type consensus sequences are present at the appropriate positions 5' upstream from these putative transcriptional initiation sites for $MLC1_F$ and $MLC3_F$. We therefore conclude that the two MLC mRNAs are probably generated by distinct transcription initiation events, followed by alternative splicing (1).

The functional role of the alkali myosin light chain isoforms $MLC1_F$ and $MLC3_F$ is not yet evident, which makes it more difficult to propose a rationale for the evolution of such a gene. In adult fast skeletal muscle, these two proteins are present in approximately similar amounts, for example in rabbit they have been shown to be in the molar ratio of 1 : 0.85 (7,8) but they do not co-accumulate at all developmental stages : $MLC1_F$ is present in foetal skeletal muscle, while $MLC3_F$ only accumulates later (9,10,11). This is confirmed at the mRNA level in mouse foetal skeletal muscle (12). The structure of the alkali myosin light chain gene has also been described for the chicken (3) and the rat (2). We have undertaken a sequence comparison of the genes from mouse, rat and chicken. There is a striking conservation of blocks of sequences located 5' upstream from the Cap sites of $MLC1_F$ and $MLC3_F$, even between species as remote as chicken and mouse, which would suggest a functional requirement for these sequences. Other features of the gene, which have been conserved during evolution, may have a role in the complex mechanism of differential splicing which is required for the production of two mRNAs from this locus.

MATERIALS AND METHODS

Computer analysis -Nucleotide sequences have been compared using a self-serve sequence analysis system implemented on a Data General MV8000 and a Nova 3 computer at the Institut Pasteur (I.P). The printing of sequences in parallel and the calculation of percentage homology in sequences have been performed using the PRTPARAL.F77 program worked out by O.Gérard (I.P). A search for similarity between a nucleotide probe and any other nucleotide sequence has been done using the SEQFIT.F77 program adapted by B. Caudron (I.P) from R. Staden (M.R.C, Cambridge). Research of homologies,

palindromic or inverted repeats was performed using a dot matrix comparison program for DNA sequences : it compares each set of 20 bp. of a sequence, placed on the ordinate, with each set of 20 bp. of another sequence, placed on the abscissa, and draws a dash when two sets display 70 % homology or more. Codon usage within a coding sequence, from which the bias values were calculated, was established using a program adapted by J.M Claverie (I.P) from that of Staden and McLachlan (13).

DNA sequencing -Sequences of the mouse promoter regions of $MLC1_F$ and $MLC3_F$ were done as previously described (1). Fragments were generated at random from total recombinant phage DNA (10 micrograms) by sonication, then end repaired with DNA polymerase I, Klenow fragment (Boehringer, Mannheim) as proposed by Messing (14). Fragments 500 to 700 bp. in size were selected on agarose gels and ligated into the Sma I site of M13mp8 (15). Recombinants corresponding to 5' flanking sequences of the promoters and of the first specific exons of $MLC1_F$ and $MLC3_F$ were sequenced by the dideoxy method (16,17).

Mapping the Cap sites of mouse $MLC3_F$ by primer extension experiments

Primer extension was performed essentially as described in (18). Labelled probes were prepared from a recombinant M13-mp8 template covering the promoter region and the first exon of $MLC3_F$ mRNA, by primer extension from the universal 15 bp M13 primer (Biolabs, New England). Suitable single stranded regions were purified following restriction and gel electrophoresis. 5000 cpm (Cerenkov) of probe were hybridized to 17 µg of total RNA from new-born mouse muscle, for 3 hours at 60°C, in 10 µl of 0.1 M NaCl, 20 mM Tris (pH 8.3) and 0.1 mM EDTA. After dilution with 2 x concentrated transcription buffer, primed molecules were extended with 7 units of AMV reverse transcriptase (Genofit) and unlabelled deoxynucleotides. Elongated products were analysed on 6% acrylamide sequencing gels.

RESULTS AND DISCUSSION

STRUCTURAL COMPARISON OF MOUSE, RAT AND CHICKEN $MLC1_F/MLC3_F$ GENES

The gross structure of the $MLC1_F/MLC3_F$ gene is very similar in mouse (1), chicken (3) or rat (2). As shown in figure 1, the common carboxy-terminal part of the two proteins is encoded by four exons (numbered 5 to 8), the specific NH_2-terminal ends are specified by four other exons, two for $MLC1_F$ (exons 1 and 4) and two for $MLC3_F$ (exons 2 and 3). In both the mouse and chicken genes there is an intervening sequence in

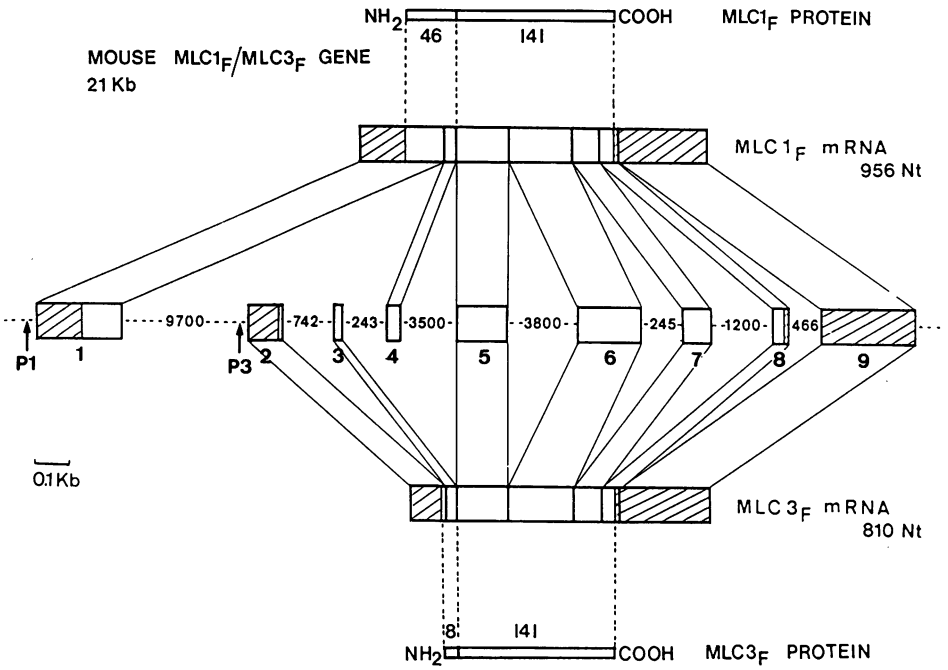


Figure 1. Schematic representation of the $MLC1_F/MLC3_F$ exon and intron organization in the mouse genome. The size of the introns is indicated in base pairs and the exons are drawn to scale. Exons are numbered 1 to 9, from 5' to 3'. Dashed boxes represent non coding regions of the mRNAs. The size of the mRNAs are indicated in nucleotides (Nt); they have been calculated from the gene sequence itself and the length of the poly A tail is not taken into account here. Numbers relative to the proteins represent the number of aminoacid residues in specific NH_2 terminal parts or common COOH terminal ends. Arrows indicate positions of the promoter regions for $MLC1_F$ (P1) and $MLC3_F$ (P3).

the 3' non coding region, which is bordered by correct donor and acceptor splice signals. Information on the position of this intron in the rat gene is not yet available. Such an intron is also found in the 3' non coding region of the MLC-ALK gene in drosophila (19).

The total length of the gene, from the $MLC1_F$ cap site to the polyadenylation signal is very similar: the chicken gene (17.6 Kb) is shorter than that of the mouse (21 Kb) or rat (21.5 Kb). The ratio of total exon/intron length is exceptionally high (1/18), compared to that of most other eucaryotic genes .

Introns and splicing. The length of introns is very variable. It ranges from 245 nucleotides (average size for intron 3) to 10 kilobases (average size for intron 1). Comparison of the length of individual introns reveals

Table 1. Comparison of the length of introns and exons between mouse, rat and chicken $MLC1_F/MLC3_F$ genes. The sizes of introns and exons are indicated in nucleotides, and percentages of homology in a given exon between mouse and chicken, mouse and rat or rat and chicken genes have been calculated by computer analysis. Abbreviations are C for Coding, NC for Non Coding and N D for Not Determinated.

	SIZE IN NUCLEOTIDES			% HOMOLOGY		
	MOUSE	RAT	CHICK	MOUSE /CHICK	MOUSE /RAT	RAT /CHICK
5'NC	126	122	131	51	92	47
EXON 1	239	245	257			
C	114	123	126	63	89	64
INTRON 1	9700	11000	9200			
5'NC	94	83	71	36	58	37
EXON 2	97	86	74			
C	3	3	3	100	100	100
INTRON 2	742	800	577			
EXON 3	25	25	25	60	100	60
INTRON 3	243	244	246			
EXON 4	28	28	28	89	96	85
INTRON 4	3500	4000	3150			
EXON 5	144	144	144	76	93	74
INTRON 5	3800	2900	1514			
EXON 6	174	174	174	80	98	81
INTRON 6	245	250	467			
EXON 7	78	78	78	78	96	79
INTRON 7	1200	1100	503			
C	26	26	26	88	96	89
EXON 8	43		53			
3'NC	17	285(total)	27			
INTRON 8	466	ND	804			
EXON 9 (3' NC)	250	ND	318	30	83 (total 3' NC)	30
TOTAL LENGTH	21000	21500	17600			
GLOBAL HOMOLOGIES OF CODING REGIONS				79	96	79

that the size of the four introns (1 to 4) located in the isoform specific 5' portion of the $MLC1_F/MLC3_F$ gene is conserved between the three species, whereas intron size differs more between species in the 3' portion of the gene containing the common exons (see table 1). The average values for intron size variation are about 10 % for introns 1 to 4, and 52 % for introns 5 to 8 . This particularity is well illustrated by the case of intron 3 which has the same size, to within 3 nucleotides, in the mouse, rat and chicken genes. Size conservation of the first large intron may reflect a requirement for maintaining a minimal distance between the two promoters.

Intervening sequences are located at exactly the same position in the coding sequences in these three genes. The type of intron-exon junction at the 5' border of exons 3 and 4, specific for $MLC1_F$ and $MLC3_F$ respectively, differs from that of all the other common exons 5 to 8: the joining of the first two NH_2-terminal specific exons for either $MLC1_F$ or $MLC3_F$ always occurs between two codons whereas any splicing event involving a common exon occurs between the first and the second nucleotide of a codon. Thus, either for $MLC1_F$ or $MLC3_F$, a first specific exon has to be joined to a second specific exon before being spliced to a common exon, in order to retain the same reading frame.

Where splice junction sequences are concerned, the donor sequences at exon-intron boundaries (5' splice sites) and the acceptor sequences at intron-exon boundaries (3' splice sites) are in good agreement with the consensus sequences proposed by Mount (20),(see table 2). Specific exons 3 and 4 have exactly the same donor site GTAAGT at their 3' border which in all three genes is then joined to the same acceptor site CTTGCAG at the 5' border of the first common exon. This may reflect a more specific requirement for a given splice junction sequence around intron 4 which can be regarded as a "hinge" point in the gene. In the mouse $MLC1_F/MLC3_F$ gene, a consensus sequence Py.Py.Pu.A.Py is located between 14 and 47 nucleotides (with an average of 28) upstream from the 3' splice site in all introns, in accordance with the proposal that such a sequence represents a putative branch point for excision of introns and is a common feature of mammalian intervening sequences (21).

Comparative examination of the sequence and structure of the $MLC1_F/MLC3_F$ gene gives some indication as to how an alternative splicing may occur for the generation of functional $MLC1_F$ and $MLC3_F$ mRNAs. The way in which the exons are joined together, especially in the 5' portion of the

Table 2. Comparison of splice junction sequences in rat, mouse and chicken $MLC1_F/MLC3_F$ genes. For each donor or acceptor splice site, rat, mouse and chicken sequences were obtained from published data (2,1,3) and written on top of each other as indicated for exon 1. The symbol / is used for delimiting the junction between exons and introns. N D is the abbreviation for Not Determinated. Consensus sequences are from Mount (20).

	5' SPLICE SITE (donor site)		INTRON	3'SPLICE SITE (acceptor site)	
EXON 1	G/GTAACT G/GTAACT G/GTAATG	(rat) (mouse) (chicken)	1		
EXON 2	G/GTGGTC G/GTGGGTC G/GTGGGTT		2	GCTCCAG/T ACCACAG/T ACTGCAG/T	EXON 3
EXON 3	G/GTAAGT G/GTAAGTA G/GTAAGTA		3	CCGACAG/A CCAACAG/A CGACCAG/A	EXON 4
EXON 4	G/GTAAGT G/GTAAGTT G/GTAAGTT		4	CTTGCAG/A CTTGCAG/A CTTGCAG/A	EXON 5
EXON 5	G/GTAGGT G/GTAGGTTT G/GTAAGAAC		5	CTTCCAG/A CTTCCAG/A TTTGCAG/A	EXON 6
EXON 6	G/GTAAGG G/GTAAGGG G/GTAAGGG		6	TCCCCAG/G TCCCCAG/G CCTCCAG/G	EXON 7
EXON 7	G/GTACAG G/GTACAGC G/GTACGTG		7	CCAACAG/C CCAACAG/C TCCGCAG/C	EXON 8
EXON 8	N.D G/GTAGACA G/GTACGTT		8	N.D CCCGCAG/C TTCCTAG/A	EXON 9
	A G/GT AGT G	CONSENSUS SEQUENCE (Mount , 1982)		C C ( )$_n$N AG/G T T	

gene, is not in keeping with a simple linear scanning model for splicing, in which a donor site will be spliced to the most proximal acceptor site, whether this proceeds from 5' to 3' or indeed from 3' to 5'. It seems probable that the choice of the promoter and the nature of the primary transcript determine subsequent splicing events for the $MLC1_F/MLC3_F$ gene.

When initiation occurs at the $MLC1_F$ promoter, the donor splice site of

exon 1 can potentialy either be joined to the acceptor site of exon 3 or 4. In theory, a "pseudo $MLC1_F$ transcript" can be generated by joining exon 1 ($MLC1_F$ specific) to exon 3 ($MLC3_F$ specific), but there is no evidence for the accumulation of such a modified $MLC1_F$ protein. For correct splicing of exon 1 to exon 4, the secondary and tertiary structures of the transcript may play an important role in influencing the relative position of a pair of 5' and 3' splice regions, as proposed by the model for splice site selection (22). The striking conservation of intron sizes in the 5' region of all three $MLC1_F/MLC3_F$ genes, and especially of intron 3 as discussed previously, may be related to structural requirements for the primary transcripts, related to the alternative splicing mechanism. Morover, there is a direct sequence homology within intron 1 at 52 nucleotides from the donor site of exon 1, and within intron 3, at 67 nucleotides from the acceptor site of exon 4 in the mouse gene : AGAACTC(A/G)AGGATT---TGGA . This may be significant either as a splicing signal or in the formation of an appropriate secondary structure, in the promotion of the correct exon 1/exon 4 joining for the $MLC1_F$ mRNA. No such direct sequence homology can be found in common with intron 1 and intron 2.

In the case of initiation at the $MLC3_F$ promoter, examination of the nucleotide sequences provides no obvious indication of why splicing occurs between exons 2 and 3 and not between exons 2 and 4. The secondary structure of the $MLC3_F$ primary transcript may again provide an explanation. There is a 23 nucleotides long sequence, composed only of pyrimidine residues and located at a distance of 10 nucleotides upstream from the acceptor splice site of intron 2, which is 75% homologous between the mouse and the chicken $MLC1_F/MLC3_F$ genes. Such a similarity, in a region essential for the excision of intervening sequences of pre-mRNAs (21,23), is unique when all introns are compared between these two species and so may be involved in the selection of the neighbouring splice site and the alternative splicing mechanism. Preference for one acceptor splice site may be imparted by factors or proteins that change the spatial structure of the $MLC3_F$ pre-mRNA or cover the splice regions. As previously mentioned, the intron-exon boundary sequences of the specific exons 3 and 4 are such that their combination would generate a one base frameshift, resulting in a nonsense codon TAA in exon 4 in the mouse or rat gene, and a TGA stop codon at the junction of exons 4 and 5 in the chicken gene.

Exons : coding and non coding regions.. The nucleotide length of each coding exon is conserved between mouse, rat and chicken, except for exon 1

(specific for $MLC1_F$). Indeed comparison of aminoacid sequences for rabbit (4) and chicken (6) demonstrates that the two alkali myosin light chain proteins are very conserved in length and sequence between these two species. It was therefore surprising to find that whereas exon 1 in the chicken gene codes for the 41 aminoacids described in the chicken $MLC1_F$ protein, it only encodes 37 aminoacids in the mouse and 40 in the rat instead of the 41 aminoacids of rabbit $MLC1_F$ (4), the initiation codon for methionine is not taken into account in these figures. In both cases, a deletion has occurred in the Ala-Pro-Ala rich NH_2-terminal region of the protein. The nucleotide sequence in this region is very GC rich and it is possible that there has been a deletion at this position during propagation of the recombinant lambda phage , although the fact that it is seen in two independent experiments makes this less probable. Alternatively, this sequence may be subject to variation and deletions may have occured in this sequence of the gene in the course of mammalian evolution. As seen in table 1, the nucleotide sequence of the coding regions is very conserved between mouse and rat (96 % homology) whereas it is more divergent between chicken and either mouse or rat (79 % homology).

When the codon usage in the $MLC1_F/MLC3_F$ gene is examined by calculating the "bias value" (G+C/A+T) for the third nucleotide position within a given codon, very similar values are found for the mouse (1.60),the rat (1.50) and the chicken gene (1.53). Comparison of codon usage in the skeletal α actin and fast myosin light chain mRNAs, which are both major species expressed in adult skeletal muscle, shows that this is quite different as indicated by the different bias figure : 3.4 for the mouse and rat and 2.7 for the chicken skeletal muscle α actin gene (24). A simple model which correlates codon usage and the composition of the tRNA pool for abundant mRNA species in a given tissue, is thus excluded in this case.

In contrast to the coding regions, the 3' untranslated regions common to $MLC1_F$ and $MLC3_F$ mRNAs are less conserved between species; mouse and rat sequences are 83 % homologous, while the chicken sequence is only 30 % homologous to that of the rat or mouse gene, a value equivalent to that expected between any random sequences (table 1). These values are the same whether the 5' or 3' portions of the 3' untranslated regions of these genes are compared. This is in contrast to findings of Miyata et al.(25), which showed that the distal parts of some 3' non coding regions are more conserved than the proximal parts. A common sequence element,

G(A/T)AGACTGG(A/C)CA, is found in the proximal region of the 3' untranslated sequence of mouse and rat, located 29 nucleotides after the TAA stop codon, and is also present in the 3' untranslated region of the chicken gene, but further downstream at 120 nucleotides from the stop signal. In the case of the rat and mouse gene, this sequence is flanked by two direct repeats TTCAAGAA or TTCAAG, absent in the chicken gene. It has been proposed for the rat (2) that this is a repetitive element (see also 1), common to a tissue specific family of transcripts. Another sequence element is also found in the mouse and rat but not in the chicken $MLC1_F/MLC3_F$ gene at 167 and 180 nucleotides respectively after the stop codon. This same sequence, TCAGGA(T/C)GACAATC, is found 30 nucleotides after the stop codon in the 3' untranslated region of the α skeletal actin gene of chicken and rat (26).

The 5' non coding sequences of $MLC1_F$ mRNAs show 92 % homology between mouse and rat, and there is about 50 % divergence in this region between either mouse and chicken, or rat and chicken (see table 1). This higher figure for the 5' non coding region of avian and mammalian $MLC1_F$ mRNAs is mainly due to sequence homology in the more 5' distal part of the sequence (see figure 2). Comparison of 5' non coding sequences between $MLC3_F$ mRNAs reveals that they are much more diverged between these species than the 5' non coding $MLC1_F$ sequences. Unexpectedly only 58 % homology is detected between mouse and rat , and these sequences are completely diverged (36 and 37 % homology) between these mammalian and avian genes.

COMPARATIVE STUDIES ON PROMOTER REGIONS OF THE $MLC1_F/MLC3_F$ GENE

The regions in the mouse $MLC1_F/MLC3_F$ gene, upstream from the specific NH_2 terminal exons of $MLC1_F$ (exon 1) and $MLC3_F$ (exon 2) have now been extensively sequenced (see figures 2 and 4). Comparison with the chicken gene and, when the data is available, with the rat, clearly points to blocks of conserved sequences, interspersed with dissimilar sequences, in a region which, like most other flanking and intron sequences, is otherwise highly diverged between birds and mammals.

The promoter region for $MLC1_F$. The sequence comparison of $MLC1_F$ promoter regions reveals several blocks of nucleotides which are highly conserved between mouse, rat and chicken (figure 2). In fact the whole region from -145 to +30 is conserved (75 % homology), the presence of short sequence elements showing more divergence led us to describe a series of blocks.

A first block of homology is 97 nucleotides long and includes the

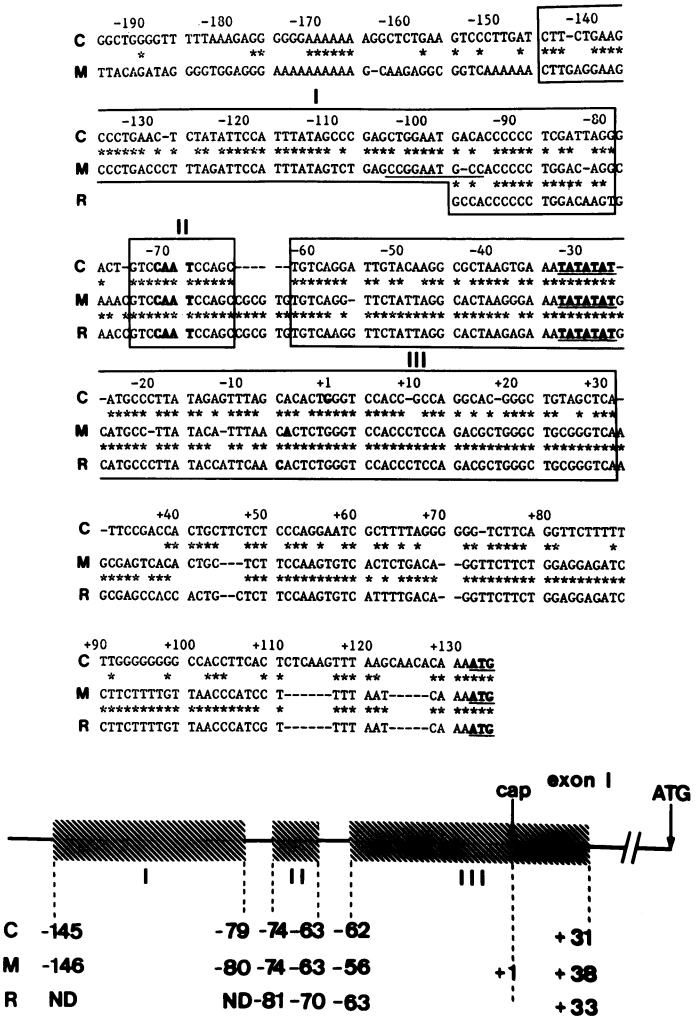


Figure 2. Comparison of $MLC1_F$ promoter region sequences in chicken, mouse and rat $MLC1_F/MLC3_F$ genes. The boxed regions marked I, II and III represent regions of strong homology (more than 70 %) within all three (where data available) chicken (C), mouse (M) and rat (R) $MLC1_F$ promoters. Numbers represent the nucleotide position relative to the chicken $MLC1_F$Cap site (noted by +1). Single nucleotides in bold type indicate positions of transcription initiation determinated by S1 protection or primer extension experiments. Asterisks indicates matches, and dashes are for gaps introduced for optimal alignment. CCAAT, ATA consensus sequences, initiation codons ATG and putative enhancer core sequences are underlined. On the lower schematic representation, dashed boxes represent the same regions of homology as those shown on the upper sequences, and numbers indicate their arbitrary limits relative to their corresponding Cap sites (noted by +1). N D is for Not Determinated.

$MLC1_F$ putative Cap sites and ATA box regions for the three genes (region III from -56 to +38 in the mouse sequence, -63 to +33 in the rat sequence and -62 to +31 in the chicken). This region , comprising the transcription start signals , is 92 % homologous between rat and mouse, and 78 or 76 % between mouse/chicken and rat/chicken respectively. All three $MLC1_F$ promoters have the same sequence : AAATATATAT as an ATA-like consensus sequence, similar to that of most eucaryotic RNA polymerase II promoters of protein coding genes (27).

A second region of homology is a short sequence of 12 nucleotides comprising the CCAAT box (28) which is perfectly conserved between the three genes (region II :GTCCAATCCAGC). Such a strong sequence conservation in the CCAAT region has also been noted between chicken and rat for the skeletal muscle α actin gene (26) and between chicken and mouse for the $\alpha 2$(I) collagen gene (29). Distances in nucleotides between the CCAAT and ATA boxes are similar in these $MLC1_F$ promoter regions: 35 nucleotides in the chicken gene and 40 or 41 nucleotides in the mouse or rat genes .

A third block of homology is located 5' upstream from the CCAAT box, between -80 and -146 in the mouse sequence, and between -79 and -145 in the chicken $MLC1_F$ promoter (see region I in figure 2). This 69 nucleotide long sequence is of particular interest because it shows 76 % homology between the two genes: this value is of the same order as the 79 % homology noted for $MLC1_F/MLC3_F$ coding sequences (on which a functional pressure must be exerted during evolution), and much higher than the 51 % ($MLC1_F$) and 36 % ($MLC3_F$) values noted for the 5' and 30 % value noted for the 3' non coding regions of the mRNAs (see table 1). Region I represents one of the first examples described of a relatively long sequence upstream from the CCAAT box which is highly conserved between mouse and chicken, two species which are separated by an evolutionary distance of 250 million years (30). Conserved but shorter sequences including the ATA box, the CCAAT box and a G/C rich sequence located around position -110 have been reported for the mouse and chicken $\alpha 2$(I) collagen genes (29).

In addition to regions I, II and III, A/G rich sequences are located further 5' upstream at about the same distances from the start of transcription in both mouse and chicken $MLC1_F$ promoter regions (-165 to -192 and -163 to -183 respectively).

The fact that the nucleotide sequences of region II (including the CCAAT box) and of region III (including the ATA box) are so highly conserved between rat, mouse and chicken genes, and that region I is very

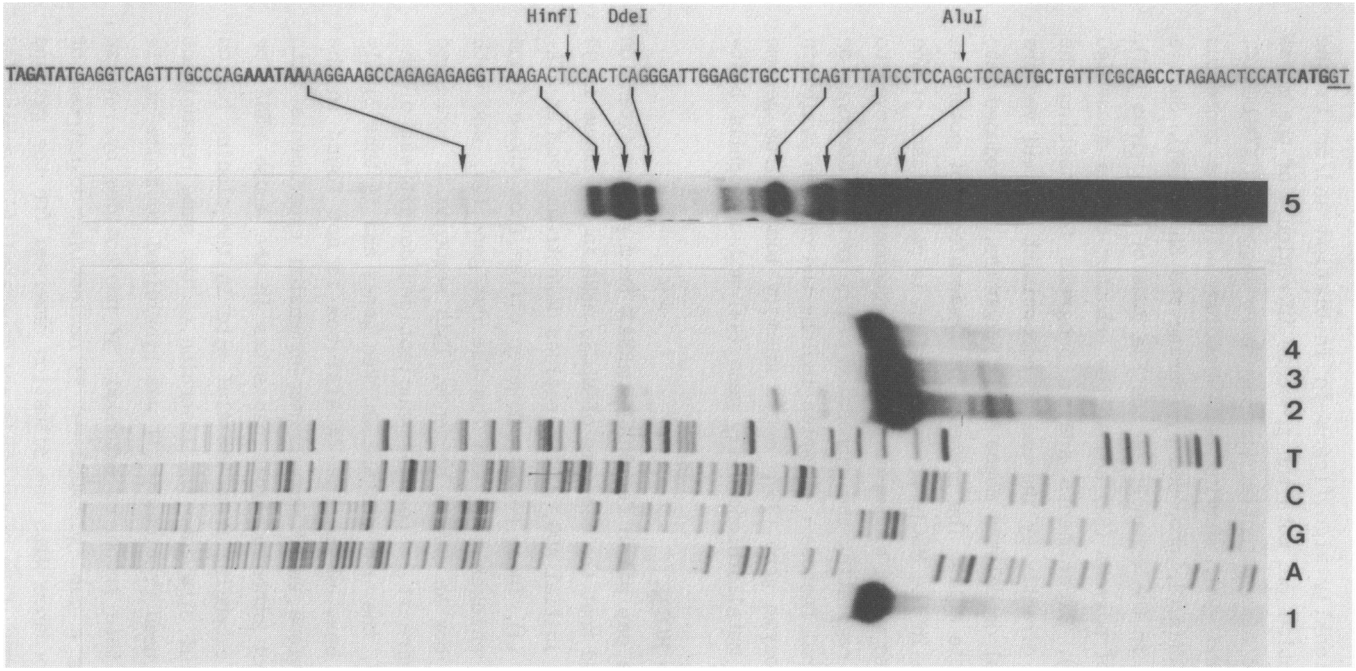
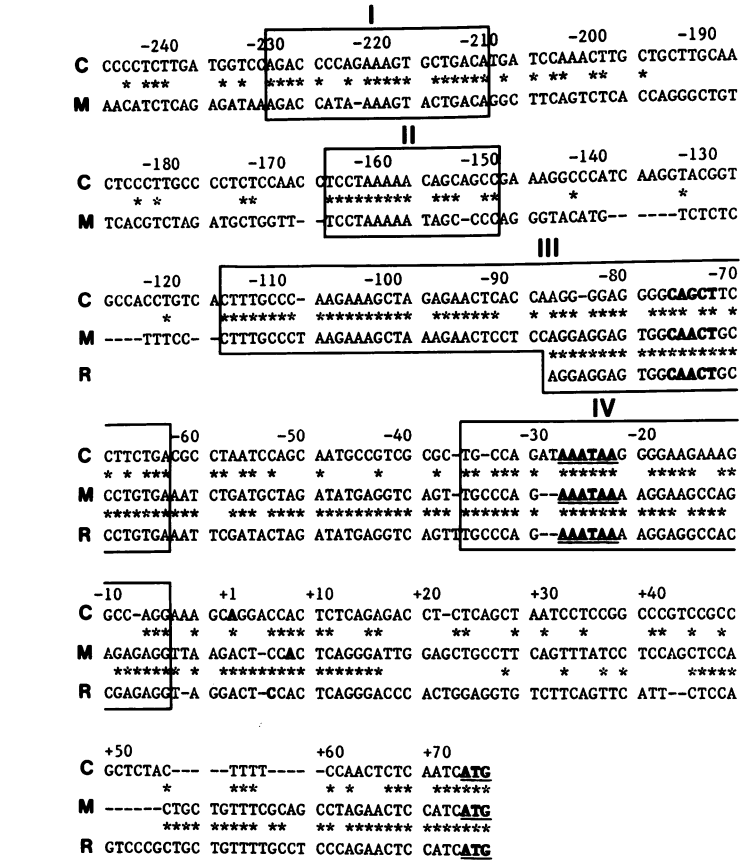


Figure 3. Mapping of multiple cap sites for $MLC3_F$ mRNA by primer extension analysis. A 105 nucleotide fragment extending up to the AluI restriction site shown in the sequence was used to prime reverse transcription on mouse muscle RNA. Resulting fragments were analysed on sequencing gels (lane 2). The sequencing products of a recombinant M13-mp8 were comigrated as size markers (lanes A, G, C and T). Lane 1 and 4: the non elongated probe; lane 3: the same probe incubated with E. Coli tRNA instead of mouse muscle RNA prior to reverse transcription; lane 5: same as 2, but exposed for 6 days instead of 15 hours. On the sequence are indicated the DdeI and HinfI sites which were used in other primer extension experiments (data not shown; and 1). The short arrows point to the end of the fragments generated on the other strand (coding strand) by these restriction enzymes.

homologous between the mouse and chicken sequences would strongly suggest that they are functionally significant in initiation and/or in the regulation of transcription. We note the presence in region I of the mouse $MLC1_F$ promoter sequence of a potential enhancer core sequence (CCGGAATGCC, located between -107 and -94), showing 82 % homology with the consensus sequence , C(A/G)GGAAGTGA(A/C), proposed by Hearing and Shenk (31). Such a core sequence is found in the enhancer of the E1a gene of adenovirus (31), in the major enhancer of polyoma (32) and in the enhancer of the mouse IgG heavy chain gene (33). Experiments are in progress for testing the importance of region I directly by the use of recombinant plasmid vectors containing either normal or deleted mouse $MLC1_F$ promoters, introduced into differenting mouse muscle cells. The presence of these blocks of potential regulatory significance in the $MLC1_F$ promoter region is reminiscent of the situation for other genes where blocks of promoter sequences have proved to be important either in regulating the expression of the gene by hormones or in binding transcriptional factors : examples include the chicken ovalbumin gene (34), chicken lysozyme gene (35), mouse mammary tumor virus genome (36,37,38,39,40), human metallothionein II_A gene (41) and drosophila heat-shock genes (42,43).

The promoter region for $MLC3_F$.

Evidence for multiple cap sites for $MLC3_F$ mRNA.. From primer-extension experiments, we proposed that the cap site of $MLC3_F$ mRNA in the mouse is 94 nucleotides 5' from the initial AUG, 33 nucleotides downstream from an ATA-like sequence (TAGATAT) (1). These results however differ from those found previously (3) on the chicken gene and on the rat gene (2), who both propose as the Hogness box an AAATAA sequence. This sequence is conserved between the mouse (where it is found 23 nucleotides downstream of the TAGATAT box (see fig.4)), the rat and the chicken genes and is thus likely to be the functional box, while the TAGATAT box is not found in the chicken. We have therefore undertaken primer-extension experiments with shorter probes to detect other possible cap sites in the mouse mRNA for $MLC3_F$. A probe (fig.3) extending from the AluI site in the cloning region of M13 to the first AluI site into the insertion of the same recombinant M13-mp8 phage that we used previously (1) was hybridised to mouse muscle RNA. Fig.3 shows that this probe is extended 37 nucleotides by reverse transcription, defining a cap (ACTCAGGG---) 33 nucleotides downstream of the AAATAA box, in accordance with the situation observed for the rat and the chicken genes. However, five other bands are generated, all terminating



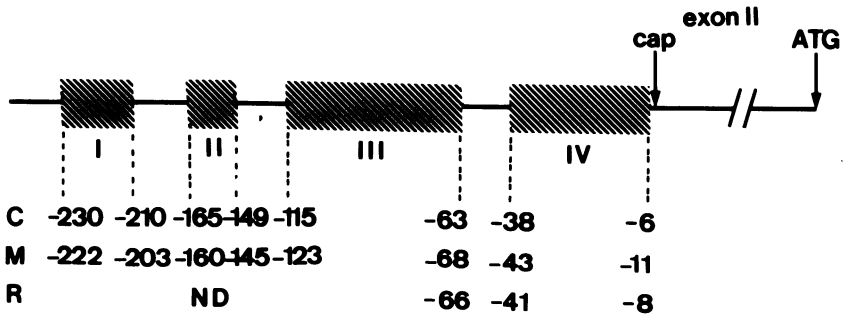


Figure 4. Comparison of the MLC3$_F$ promoter region sequences in chicken, mouse and rat MLC1$_F$/MLC3$_F$ genes. This figure is constructed exactly as for figure 2. Single boxed nucleotides represent the start of mRNA transcription determinated either by S1 protection or primer extension experiments. This corresponds to the most frequently used position detected for the mouse (see figure 3).

at an A (which is the preferential nucleotide for initiation in eucaryotes, see 27) in the sequence, which we interpret as secondary cap sites. The abundance of these other bands is not much lower than that of the main band, except for the longer fragment, which is detected only after long exposure of the gel. This corresponds to the cap site generated from the TAGATAT box, which is therefore probably less frequently used for initiation. Results obtained with the AluI probe, with a probe terminating at an upstream DdeI site (data not shown) and with one terminating at an upstream HinfI site (1) are in agreement (for definition of probes, see figure 3); the extended fragment from the HinfI restriction site which led to detection of the most 5' cap site, was previously (1) underestimated by 4 nucleotides due to an anomalous migration of the sample on sequencing gels. We therefore number the sequence (fig.4) from the cap site ACTCA---. The detection of multiple cap sites for $MLC3_F$ mRNA is substantiated by observations of Periasamy et al.(2), who define a major cap site by S1 mapping at a position similar to that now reported here, but have cloned a cDNA which extends further upstream of this site. Many eucaryotic genes have been shown to exhibit such microheterogeneity at the cap site (reviewed by Manley (44); see also 45). The molecular mechanisms responsable for this phenomenon are not clear. It may be correlated, in the case of the $MLC3_F$ promoter, to the relative divergence of the AAATAA box from the canonical TATA sequence. Around position -80 relatively to the main cap site of the $MLC3_F$ mRNA, a GGCAACT sequence is found, which is highly conserved between mouse, rat and chicken (where it reads GGCAGCT). This sequence might fulfil the role of a CAAT box (46); it is very similar to the CAAT sequence (GACAACT) proposed for the chicken lysozyme gene (47).

Sequence conservation between species.. The same type of sequence comparison, as that effected for $MLC1_F$, in the $MLC3_F$ promoter regions located at the 3' border of intron 1 and 5' upstream from exon 2, reveals four blocks of homology between mouse and chicken $MLC1_F/MLC3_F$ genes, interspersed with non homologous sequences. In this case, distinct blocks of homology are more evident than for the $MLC1_F$ promoter region. The first block (see figure 4, region IV) is 36 nucleotides long and lies between positions -11 to -43 , -6 to -38 and -8 to -41 in the mouse , chicken and rat sequences recpectively, showing 70 % homology. This region comprises the ATA consensus sequence (AAATAA) which is found at -33, -30 and -27 nucleotides from the corresponding cap sites in the mouse, rat and chicken genes. A second and much longer sequence of 56 nucleotides, region III ,in

the mouse (from -68 to -123) is 76 % similar to a 54 nucleotide block in the chicken gene (-63 to -115). Less sequence data are available for the rat $MLC3_F$ promoter region, but the rat and mouse sequences 5' further upstream from the putative $MLC3_F$ Cap sites are very conserved (90 % homology), so that region III, homologous between mouse and chicken, is also found in the rat gene for the part sequenced . This region comprises the CAACT sequence as mentioned before.

5' further upstream, two other blocks of homology (region II and I on figure 4) can be delimited : they both show 80 % of homology between the two species and are 16 or 17 nucleotides long respectively at positions -145 /-160 in the mouse and -150 /-167 in the chicken, and 20 or 21 nucleotides at positions -203 /-222 (mouse) and -211 /-231 (chicken). The $MLC3_F$ promoter region therefore ressembles the $MLC1_F$ promoter in that multiple boxes of conserved sequences, with 70 to 100 % homology separated by dissimilar sequences, are present in the three species. One important point in these comparisons is that, within the same species, there is no very clear evidence of sequence homology between the $MLC1_F$ and $MLC3_F$ promoters, which might be expected if a common transcriptional factor activates the transcription of both mRNAs. The only homology found in the chicken MLC promoters is a 15 nucleotide long sequence CCTCGATT(A/-)GGGACT at position -75/-89 in the $MLC1_F$ promoter and -306/-320 in the $MLC3_F$ promoter. A somewhat different sequence of 14 nucleotides, located at similar distances from the respective cap sites is present in the mouse gene (AG(A/-)TTCCATT(A/T)ATA) at -111/-124 and -362/-374 respectively in the $MLC1_F$ and $MLC3_F$ promoters. These are the only sequences that can be detected as 80/90 % homologous in the 5' flanking sequences of either mouse or chicken $MLC1_F$ and $MLC3_F$ promoters. They do not fall within the blocks of 5' upstream sequences conserved between species, and again deletion experiments should show whether these species specific common $MLC1_F/MLC3_F$ sequence elements have any regulatory significance.

COMPARISON OF PROMOTER SEQUENCES BETWEEN DIFFERENT CONTRACTILE PROTEIN GENES

The general question of whether genes expressed in the same cell phenotype have common promoter sequence elements which may be recognized by a "phenotypic" transcriptional factor, can now be asked for muscle where the necessary structural information is available for genes encoding the skeletal muscle α actin (48), a regulatory myosin light chain $MLC2_F$ (49),

and the alkali myosin light chains $MLC1_F$ and $MLC3_F$ (2) co-expressed in adult fast skeletal muscle fibres of rat, and chicken (50,3). No common sequence element is found at the same position from the cap site in mouse, rat or chicken, in the 5' promoter regions of $MLC1_F$, $MLC3_F$ or skeletal muscle α actin genes. In the rat, examination of promoter regions reveals that a short 11 nucleotide sequence GGG(C/T)A(A/G)GG(C/T)(C/T)A is present both in the skeletal muscle α actin gene at -186, and in the $MLC2_F$ gene at -64. This sequence is also found in the promoter region of the rat cardiac myosin heavy chain gene (51) at -200, and may therefore not be related to co-expression of these genes since, although some transcripts are detected from the α skeletal actin gene (52), there is no evidence for transcription of $MLC2_F$ in the heart. Similar comments apply to a 20 nucleotide sequence (CCCTGACCCTTTAGATTCCA) located between -121 and -140 in the mouse $MLC1_F$ promoter region which shows 85 % homology with a sequence located between -108 and -125 in the mouse α cardiac actin promoter. A potential enhancer core sequence, similar to that noted for $MLC1_F$ around -100, is also present at the same position in the α cardiac actin gene (I. Garner, unpublished results). No common sequences are found between the $MLC1_F/MLC3_F$ and skeletal muscle α actin genes in the 5' promoter regions. The only conserved sequence element between these genes, which may therefore be significant in the coordinate expression of the α skeletal actin and $MLC1_F/MLC3_F$ genes in adult skeletal muscle, is that found in the 3' untranslated regions of the mRNAs, discussed previously. In some cases where the promoter sequences of genes expressed in the same cell have been examined, the situation is similar to that of the skeletal muscle α actin and $MLC1_F/MLC3_F$ genes: no common 5' sequence elements have been implicated for example in the co-expression of alpha and beta globin genes (53) in red blood cells. In other cases, short repeated sequences common to the 5' flanking regions of inducible genes are required for regulation and coordinate expression , as in the case of certain yeast genes for aminoacid synthesis (his 1, 3 and 4, and trp 5), drosophila heat-shock genes or glucocorticoid induced genes (reviewed in 54). Although no functional test has yet been performed, Poole and Firtel (55) suggest that short homologous G/C rich stretches found 5' to the three discoidin I genes are important in their co-expression in dictyostelium discoidum, and Fowlkes et al.(56) propose that conserved sequences in 5' flanking regions may be implicated in the regulation of transcription of coordinately expressed rat fibrinogen genes. In the case of P25 and fibroin genes co-expressed in the

silk gland of Bombyx mori, the same nucleotide sequence is clearly present in 5' flanking regions of the genes (57). It would appear that just as simple models of gene clustering can be excluded (58,59), so the presence of a common 5' sequence element binding the same transcriptional factor cannot be invoked to explain co-ordinate expression of genes in a phenotype such as that of skeletal muscle. Such sequences may lie elsewhere in the gene, or the interaction of specific and common transcriptional factors may be such that one dimensional nucleotide sequence conservation is conceptually too simple. The demonstration of tissue specific expression of a rat skeletal muscle α actin gene re-introduced into rat muscle cells (line L8) in a construct with only a limited 5' flanking region (60), suggests that the necessary sequences for the biological regulation of at least this skeletal muscle gene are present in the promoter region. Nevertheless other sequences directly implicated in the regulation of such genes may be more distant than the immediate 5' flanking regions discussed for the $MLC1_F/MLC3_F$ gene in this paper, or intragenic as for example the sequences which may be implicated in the differential expression of human globin genes in mouse erythroleukemia cells (53).

ACKNOWLEDGEMENTS The authors are grateful to P. Amati for pointing out the potential enhancer sequence in the $MLC1_F$ promoter. S. Alonso and A. Cohen provided helpful comments. We thank J.M. Claverie and his group for providing the computer programs used. This work was supported by grants to M. Buckingham from the French Ministry of Research and Technology, the Centre National de la Recherche Scientifique, the Institut National de la Santé et de la Recherche Médicale and the Muscular Dystrophy Association (MDA) of America. I. Garner is the recipient of an SERC fellowship under the Royal Society European Programme. We thank Francois Gros for support.

REFERENCES

1 .Robert, B., Daubas, P., Akimenko, M-A., Cohen, A., Garner, I., Guénet, J-L. & Buckingham, M.E. (1984). Cell, **39**, 129-140
2 .Periasamy, M., Strehler, E.E., Garfinkel, L.I., Gubits, R.M., Ruiz-Opazo, N. & Nadal-Ginard, B. (1984). J. Biol. Chem., **259**, 13595-13604
3 .Nabeshima, Y., Fujiii-Kuriyama, Y., Muramatsu, M. & Ogata, K. (1984). Nature, **308**, 333-338
4 .Frank, G. & Weeds, A.G. (1974). Eur. J. Biochem., **44**, 317-334
5 .Henry, G.D., Dalgarno, D.C., Marcus, G., Scott, M., Levine, B.A. & Trayer, I.P. (1982). FEBS Lett., **144**, 11-15
6 .Matsuda, G., Maita,T. & Umegane,T.(1981). FEBS Lett., **126**,111-113
7 .Lowey, S. & Risby, D. (1971). Nature (London), **234**, 81-85
8 .Takahashi, M. & Tonomura, Y. (1975). J. Biochem. (Tokyo), **78**, 1123-1133
9 .Sreter, F.A., Balint, M. & Gergely, J. (1975). Dev. Biol., **46**, 317-325

10.Roy, R.K., Sreter,F.A. & Sarkar,S. (1979). Dev. Biol., **69**, 15-30
11.Gauthier, G.F., Lowey, S., Benfield, P.A. & Hobbs, A.W. (1982). J. Cell Biol., **92**, 471-484
12.Barton, P.J.R., Robert, B., Fiszman, M.Y., Leader, D.P. & Buckingham, M.E. (1985). J. Muscle Res. Cell Motility, in press.
13.Staden, R. & McLachlan, A.D. (1982). Nucl. Acids Res., **10**, 141-156
14.Messing, J. (1983). in Methods in Enzymology, volume **101**, Recombinant DNA, Part C, R. Wu, L. Crossman & K. Moldave, eds. (New York : Academic Press), pp 20-78
15.Messing, J. & Vieira, J. (1982). Gene, **19**, 269-276
16.Sanger, F., Nicklen, S. & Coulson, A.R. (1977). Proc. Natl. Acad. Sci., U.S.A., **74**, 5463-5467
17.Biggin, M.D., Gibson, T.J. & Hong, G.F. (1983). Proc. Natl.Acad. Sci., U.S.A. **80**, 3963-3965
18.Panthier, J.J., Dreyfus, M., Tronik-Leroux, D. & Rougeon, F. (1984), Proc. Natl. Acad. Sci., U.S.A., **81**, 5489-5493
19.Falkental, S., Parker, V.P. & Davidson N. (1985). Proc. Natl. Acad. Sci., U.S.A.,**82**, 449-453.
20.Mount, S. (1982). Nucl. Acids Res.,**10**, 459-472.
21.Ruskin, B., Krainer, A.R., Maniatis, T. & Green, M.R. (1984). Cell, **38**, 317-331.
22.Kuhne, T., Wieringa, B., Reiser, J. & Weissmann, C. (1983). EMBO J., **2**, 727-733.
23.Keller, E.B. & Noon, W.A. (1984). Proc. Natl. Acad. Sci., U.S.A., **81**, 7417-7420.
24.Alonso, S. & Buckingham, M. (1985), submitted to J. Mol. Evol.
25.Miyata, T., Yasunaga, T. & Nishida, T. (1980). Proc. Natl. Acad. Sci., U.S.A., **77**, 7328-7332.
26.Ordahl, C.P. & Cooper, T.A. (1983). Nature, **303**, 348-349.
27.Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kédinger, C. & Chambon, P. (1980). Science, **209**, 1406-1414.
28.Efstradiatis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulder, C.C. & Proudfoot, N.J. (1980). Cell, **21**, 653-668.
29.Schmidt, A., Yamada, Y. & de Crombrugghe, B. (1984). J. Biol. Chem., **259**, 7411-7415.
30.Dickerson, R.E. (1971). J. Mol. Evol., **1**, 26-45.
31.Hearing, P. & Shenk, T. (1983). Cell, **33**, 695-703.
32.Herbomel, P., Bourachot, B. & Yaniv, M. (1984). Cell, **39**, 653-662.
33.Banerji, J., Olson, L. & Schaffner, W. (1983). Cell, **33**, 729-740.
34.Dean, D.C., Knoll, B.J., Riser, M.E. & O'Malley, B.W. (1983). Nature, **305**, 551-554.
35.Renkawitz, R., Schutz, G., Von der Ahl, D. & Beato, M. (1985). Cell, in press.
36.Payvar, F., DeFranco, D., Firestone, G., Edgar, B., Wrange, O., Okret, S., Gustafson, J.A. & Yamamoto, K.R. (1983). Cell, **35**, 381-392.
37.Scheidereit, C., Geisse, S., Westphal, H.M. & Beato, M. (1983). Nature, **304**, 749-752.
38.Buetti, E. & Diggelmann, H. (1983). EMBO J., 2, 1423-1429.
39.Hynes, N., Van Ooyen, A.J.J., Kennedy, N., Herrlich, P., Ponta, H. & Groner, B. (1983). Proc. Natl. Acad. Sci., U.S.A., **80**, 3637-3641.
40.Majors, J. & Varmus, H. (1983). Proc. Natl. Acad. Sci., U.S.A., **80**, 5866-5870.
41.Karin, M., Haslinger, A., Holtgreve, H., Richards, R.I., Krauter, P., Westphal, H.M. & Beato, M. (1984). Cell, **308**, 513-519.
42.Wu, C. (1984a). Nature, **309**, 229-234.

43.Wu, C. (1984b). Nature, **311**, 81-84.
44.Manley, J.L.(1983). Prog. Nucl. Acid Res. Mol. Biol., **30**, 195-244.
45.Selvanayagam, C.S., Tsai, S.Y., Tsai, M.J., Selvanayagam, P. & Saunders, G.F. (1984). J. Biol. Chem., **259**, 14642-14646.
46.Benoist, C., O'Hare, K., Breathnach, R. & Chambon, P. (1980). Nucl. Acid Res., **8**, 127-142.
47.Grez, M., Land, H., Giesecke, K., Schutz, G., Jung, A. & Sippel, A.E. (1981). Cell, **25**, 743-752.
48.Zakut, R., Shani, M., Givol, D., Neuman, S., Yaffé, D. & Nudel, U. (1982). Nature, **298**, 857-859.
49.Nudel, U., Calvo, J.M., Shani, M. & Levy, Z. (1984). Nucl. Acids Res., **12**, 7175-7186.
50.Fornwald, J.A., Kuncio, G., Peng, I. & Ordahl, C.P. (1982). Nucl. Acids Res., **10**, 3861-3875.
51.Madhavi, V., Chambers, A.P. & Nadal-Ginard, B. (1984). Proc. Natl. Acad. Sci., U.S.A., **81**, 2626-2630.
52.Mayer, Y., Czosnek, H., Zeelon, P.E., Yaffé, D. & Nudel, U. (1984). Nucl. Acids Res., **12**, 1087-1100.
53.Charnay, P., Treisman, R., Mellon, P., Chao, M., Axel, R. & Maniatis, T. (1984). Cell, **38**, 251-263.
54.Davidson, E.H., Jacobs, H.T. & Britten, R. (1983). Nature, **301**, 468-470.
55.Poole, S.J. & Firtel, R.A. (1984). J. Mol. Biol., **172**, 203-220.
56.Fowlkes, D.M., Mullis, N.T., Comeau, C.M. & Crabtree, G.R. (1984). Proc. Natl. Acad. Sci., U.S.A., **81**, 2313-2316.
57.Couble, P., Chevillard, M., Ravel-Chapino, P. & Prudhomme, J.C. (1985). Nucl. Acid Res., **13**, 1801-1814.
58.Czosnek, H., Nudel, U., Mayer, Y., Barker, P.E., Pravtcheva, D.O., Ruddle, F.H. & Yaffé, D. (1983). EMBO J., **2**, 1977-1979.
59.Robert, B., Barton, P., Minty, A., Daubas, P., Weydert, A., Bonhomme, F., Catalan, J., Chazottes, D., Guénet, J-L. & Buckingham, M.E. (1985). Nature,**314**,181-183.
60.Melloul, D., Aloni, B., Calvo, J., Yaffé, D. & Nudel, U. (1984). EMBO J., **3**, 983-990.