Brader, M. L., Kaarsholm, N. C., & Dunn, M. F. (1990) *J. Biol. Chem. 265*, 15666-15670.
Brader, M. L., Kaarsholm, N. C., Lee, R. W.-K., & Dunn, M. F. (1991) *Biochemistry 30*, 6636-6645.
Brader, M. L., Borchardt, D., & Dunn, M. F. (1992) *J. Am. Chem. Soc.* (in press).
Brems, D. N., Brown, P. L., Nakagowa, S. H., & Tager, H. S. (1991) *J. Biol. Chem. 266*, 1611-1615.
Brill, A., & Venable, J. H. (1968) *J. Mol. Biol. 36*, 343-353.
Colacio, E., Costes, J., Kivekas, R., Laurent, J., Ruiz, J., & Sundberg, M. (1991) *Inorg. Chem. 30*, 1475-1479.
Derewenda, U., Derewenda, Z., Dodson, E. J., Dodson, G. G., Reynolds, C. D., Smith, G. D., Sparks, C., & Swensen, D. (1989) *Nature 338*, 594-596.
Dooley, D. M. (1987) *Life Chem. Rep. 5*, 91-154.
Evans, J. C., Morgan, P. H., Mahbouba, M., & Smith, H. J. (1979) *J. Inorg. Biochem. 11*, 129-137.
Fee, J. A. (1975) *Struct. Bonding (Berlin) 23*, 1-60.
Gewirth, A. A., Cohen, S. L., Schugar, H. J., & Solomon, E. I. (1987) *Inorg. Chem. 26*, 1133-1146.
Guss, J. M., & Freeman, H. C. (1983) *J. Mol. Biol. 169*, 521-563.
Guss, J. M., Merritt, E. A., Phizackerly, R. P., Hedman, B., Murata, M., Hodgson, K. O., & Freeman, H. C. (1988) *Science 241*, 806-811.
Hathaway, B. J. (1984) *Struct. Bonding (Berlin) 57*, 55-118.
Kaarsholm, N. C., Ko, H.-C., & Dunn, M. F. (1989) *Biochemistry 28*, 4427-4435.
Kitajima, N., Fujisawa, K., Moro-Oka, Y. (1990) *J. Am. Chem. Soc. 112*, 3210-3212.
Knapp, S., Keeman, T. P., Zhang, Z., Fikar, R., Potenza, J. A., & Schugar, H. J. (1990) *J. Am. Chem. Soc. 112*, 3452-3464.
Korszun, Z. R. (1987) *J. Mol. Biol. 196*, 413-419.
Messerschmidt, A., Rossi, A., Ladenstein, R., Huber, R., Bolognesi, M., Gatti, G., Marchesini, A., Petruzelli, R., & Finazzi-Agro, A. (1989) *J. Mol. Biol. 206*, 513-529.
Morpurgo, L., Rotilio, G., Agrō, A. F., & Mondovi, B. (1975) *Arch. Biochem. Biophys. 170*, 360-367.
Morpurgo, L., Rotilio, G., Agrō, A. F., & Mondovi, B. (1976) *Eur. J. Biochem. 64*, 453-457.
Palmieri, R., Lee, R. W.-K., & Dunn, M. F. (1988) *Biochemistry 27*, 3387-3397.
Peisach, J., & Blumberg, W. E. (1974) *Arch. Biochem. Biophys. 165*, 691-708.
Roy, M., Brader, M. L., Lee, R. W.-K., Kaarsholm, N. C., Hansen, J. F., & Dunn, M. F. (1989) *J. Biol. Chem. 264*, 19081-19085.
Roy, M., Lee, R. W.-K., Brange, J., & Dunn, M. F. (1990) *J. Biol. Chem. 265*, 5448-5452.
Schlichtkrull, J. (1956) *Acta Chem. Scand. 10*, 1455-1458.
Smith, G. D., Swenson, D. C., Dodson, E. J., Dodson, G. G., & Reynolds, C. D. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 7093-7097.
Solomon, E. I., Penfield, K. W., & Wilcox, D. E. (1983) *Struct. Bonding (Berlin) 53*, 1-57.
Taylor, J. S., & Coleman, J. E. (1971) *J. Biol. Chem. 246*, 7058-7067.
Thompson, J. S., Marks, T. J., & Ibers, J. A. (1979) *J. Am. Chem. Soc. 101*, 4180-4192.

Structure and Organization of the Heart Isoform Gene for Bovine Cytochrome *c* Oxidase Subunit VIIa[†,‡]

R. Sathiagana Seelan and Lawrence I. Grossman*
Department of Molecular Biology and Genetics, Wayne State University School of Medicine, Detroit, Michigan 48201
Received December 17, 1991; Revised Manuscript Received February 24, 1992

ABSTRACT: Mammalian cytochrome *c* oxidase (COX) is a 13-subunit polypeptide complex that contains 10 subunits coded by the nucleus and 3 by the mitochondria. The nuclear-encoded subunits, though of unknown function, are presumed to play a regulatory role. Three of these (subunits VIa, VIIa, and VIII) generally exist in one of two isoforms—a constitutive (L) isoform or a skeletal muscle/heart-specific (H) isoform. To study the regulation, and possibly function, of these isoforms, we have begun characterizing the genes. In this paper we describe the isolation and characterization of the gene for the bovine COX VIIa-H isoform. The gene consists of four exons spanning 1.58 kb and is associated with a CpG island. There are no canonical TATA or CCAAT boxes immediately upstream of the transcription start site. Putative DNA sequence elements associated with respiratory function, muscle gene activation, and housekeeping function are present both in the upstream regions and within introns.

Mammalian cytochrome *c* oxidase (COX)[1] is a multisubunit protein (Kadenbach & Merle, 1981; Merle & Kadenbach, 1982; Kadenbach et al., 1983) coded for by both nuclear and mitochondrial genomes. The catalytic activity of the protein has been shown to reside in the three mitochondrially encoded subunits, and little information is available on the function of the ten nuclear-encoded subunits [reviewed by Kadenbach et al. (1987) and Capaldi et al. (1987)]. These subunits are presumed to modulate COX activity, such as by binding to nucleotides, hormones, ions, second messengers, free fatty acids, or other substrates (Dowhan et al., 1985; Ka-

[†] Supported by a grant from the National Science Foundation (NSF DMB-89-05285) and the Muscular Dystrophy Association.
[‡] The nucleic acid sequence in this paper has been submitted to GenBank under Accession Number M83299.
* Address correspondence to this author.

[1] Abbreviations: nt, nucleotide; bp, base pair; COX, cytochrome *c* oxidase; COX VIIa-H and COX VIIa-L, heart and liver isoforms of cytochrome *c* oxidase subunit VIIa; *COX7aH* and *COX7aL*, corresponding isoform genes.

0006-2960/92/0431-4696$03.00/0

denbach, 1986; Malatesta et al., 1987). Although all eukaryotic organisms carry the equivalent mitochondrial subunits, the number of nuclear subunits varies with species; for instance, there are four nuclear subunits in *Dictyostelium discoideum* (Bisson et al., 1985), six in yeast (Power et al., 1984), and ten in mammals (Kadenbach et al., 1983; Takamiya et al., 1987). The number of nuclear subunits thus appears to be correlated with the evolutionary complexity of the organism (Kadenbach et al., 1988), and this implies that they have evolved for specialized functions in higher organisms.

In addition to the above complexity, three subunits, VIa, VIIa, and VIII, exist as isoforms in most mammals [reviewed by Kadenbach et al. (1987), Lomax and Grossman (1989), and Capaldi (1990)]. These isoforms have been termed the liver (L) and heart/muscle (H) types (Schlerf et al., 1988; Rizzuto et al., 1989; Lightowlers et al., 1990; Seelan & Grossman, 1991a). The H isoform is the predominant subunit in heart and muscle tissues. The L isoform, by contrast, is found in most tissues examined so far and has also been detected in the heart and muscle tissues (Van Beeumen et al., 1990; Anthony et al., 1990); it is also the exclusive isoform type present for subunit VIIa in rat (Kadenbach et al., 1989; Kennaway et al., 1990) and subunit VIII in human (Van Kuilenberg et al., 1988; Rizzuto et al., 1989).

To date, the cDNA sequences for all of the nuclear-encoded subunits in mammalian COX from at least one species have been reported, including those for the isoforms of VIa (Schlerf et al., 1988), VIIa (Seelan & Grossman, 1991a), and VIII (Lightowlers et al., 1990). The functional genes, however, have been isolated for only a few nonisoform subunits, such as cow, mouse, and rat COX IV (Bachman et al., 1987; Yamada et al., 1990; Virbasius & Scarpulla, 1990; Carter & Avadhani, 1991), mouse and human COX Vb (Basu & Avadhani, 1991; Lomax et al., 1991), human COX VIb (Taanman et al., 1991), and rat COX VIc (Suske et al., 1988); the latter three have been partially characterized. As a first step toward understanding the regulation of COX isoforms, we have sought to isolate and characterize genomic clones encoding the L and H isoforms of bovine subunit VIIa, hereafter referred to as *COX7aL* and *COX7aH*. In this paper, we present the complete nucleotide sequence of the *COX7aH* gene which, to our knowledge, represents the first mammalian tissue-specific COX gene characterized. The intron/exon organization, distribution of domains, and identification of potential regulatory motifs are described. This tissue-specific gene, interestingly, is located within a CpG island and contains no canonical TATA or CCAAT boxes in the immediate upstream region. A defect in the cognate human gene has been implicated in the pathogenesis of fatal and benign infantile mitochondrial myopathies (Tritschler et al., 1991).

Experimental Procedures

Materials. Restriction enzymes, murine moloney reverse transcriptase, and the random primer DNA labeling kit were purchased from Bethesda Research Laboratories; the GeneAmp PCR reagent kit was from Perkin-Elmer Cetus; for sequencing, the T7 polymerase kit from Pharmacia was used; [α-^{32}P]dCTP (3000 Ci/mmol), [γ-^{32}P]ATP (5000 Ci/mmol), and [α-^{35}S]dATPαS (1000 Ci/mmol) were from Amersham Corp. All other chemicals were reagent or molecular biology grade and were obtained from BRL, Fisher, or Sigma.

PCR Analysis. The polymerase chain reaction (PCR; Saiki et al., 1988) was performed in a Perkin-Elmer Cetus thermal cycler according to the protocol supplied by the manufacturer. Either 500 ng of total bovine liver DNA isolated by the method of Ciccarelli and Wetterhahn (1984) or 20 ng of cDNA was used in the reaction. The reaction was allowed to proceed for 40 cycles with denaturation at 95 °C for 1.5 min, annealing at 51 °C for 2 min, and extension at 72 °C for 3 min. The primers used were primer X [TT(T/C)GA(A/G)AA(C/T)(A/C)GIGTIGCIGA(A/G)AA, corresponding to nt 120–142 of the cDNA sequence (Seelan & Grossman, 1991a)], primer A [AT(A/G)TT(A/G)TCIGTIGCICCICC(C/T)TT, nt 208–186], primer B (CTTCTTGTGAGGGAAGGAGGCCCA, nt 296–273), and primer C (CATTTATTGATGCCTGTTCGGGTCC, nt 340–316). Primers X and A are degenerate primers made against the heart protein sequence and do not amplify the cognate liver gene sequences.

Southern Blot Analysis. Fragments were resolved on 1% standard agarose gels and blotted onto GeneScreen Plus (DuPont-New England Nuclear) according to the manufacturer's protocol. Prehybridization and hybridization were done at 60 °C in 5 × SSPE, 0.5% SDS, 2 × Denhardt's solution, 100 μg/mL salmon sperm DNA, and 1 × 10^6 cpm of randomly labeled probe (Feinberg & Vogelstein, 1983). Filters were washed first in 6 × SSPE/0.25% SDS for 30 min at room temperature, 1 × SSPE/0.5% SDS for 30 min at 37 °C, and finally in 0.2 × SSPE/0.5% SDS for 30 min at 65 °C. Washed filters were exposed to Kodak XAR-5 film with an intensifying screen at −70 °C.

Screening of Genomic Libraries. The 700-bp PCR fragment (see Results) was used as a probe to screen 500 000 plaques of a λEMBL-3 genomic library. Plaques were transferred to nitrocellulose filters and hybridized at 60 °C to the randomly labeled probe. Prehybridization and hybridization conditions were as above. Blots were washed three times in a solution containing 2 × SSC/0.2% SDS at 60 °C for 20 min and exposed to Kodak XAR-5 film with an intensifying screen at −70 °C for 6–12 h. Plaques were purified after three rounds of screening.

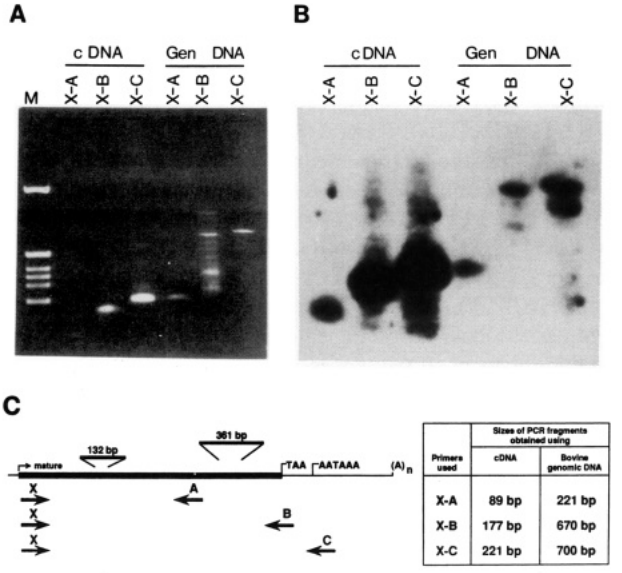
Primer Extension. Primer extension (Ausubel et al., 1989) was performed with murine moloney reverse transcriptase. A synthetic primer, complementary to the first 22 nt of the mature heart sequence, was used to prime a reaction containing 50 μg of total heart RNA. Primer-extended products were analyzed on 4% polyacrylamide sequencing gels. A sequencing reaction run in parallel was used as a size marker. After electrophoresis, the gel was fixed, transferred to GB002 paper (Schleicher and Schuell), vacuum-dried, and exposed to Kodak XAR-5 film with an intensifying screen at −70 °C for 1–10 days.

Cloning and Sequencing of Fragments. The 4.7-kb *Xho*I and the 3-kb *Eco*RI restriction fragments of λ*COX7aH1* (see Results), after isolation through low-melt agarose (Seelan & Grossman, 1991b), were ligated to *Xho*I- and *Eco*RI-digested pKS(−) vectors (Stratagene, La Jolla, CA). Deletion subclones were then generated by digestion of these clones with unique restriction enzymes present in the polylinker region and then religating the product. Additional subclones were obtained by cloning blunt-ended fragments, released by digestion of the inserts with *Alu*I, *Ssp*I, *Pvu*II, or *Rsa*I, into pKS(−) vectors digested with *Sma*I.

Appropriate aliquots of double-stranded mini-prep DNA (Mierendorf & Pfeffer, 1987) were sequenced on both strands by Sanger's dideoxy chain termination method (1977) using the M13 forward or reverse sequencing primers and sequence-specific oligonucleotides.

Results

Isolation of the Bovine COX7aH Genomic Clone. Three pairs of primers were used to amplify by PCR total bovine



Primers used	Sizes of PCR fragments obtained using: cDNA	Bovine genomic DNA
X-A	89 bp	221 bp
X-B	177 bp	670 bp
X-C	221 bp	700 bp

FIGURE 1: PCR amplification of total bovine genomic DNA. (A) PCR products (10 μL of a 100-μL reaction) were resolved on 1% agarose gels. Lane M is a *ϕX174* DNA–*Hae*III digest (marker); a control *COX7aH* cDNA or bovine genomic DNA template was used with the indicated primer pairs (X-A, X-B, X-C). (B) The gel in (A) was transferred to GeneScreen Plus and probed with labeled *COX7aH* cDNA. (C) Locations of primers on the cDNA template are shown. The start of the mature coding region, the termination codon, the polyadenylation signal, and the poly(A) tail are highlighted. The sizes of PCR fragments obtained using the cDNA and total bovine DNA as templates are depicted on the far right.

DNA in order to identify a putative intron-containing fragment. These consisted of a common forward primer, X, made against the first 23 nt of the mature heart sequence and three reverse primers, A, B, and C, located at varying distances downstream of X (Figure 1C; primer sequences and their locations are given under Experimental Procedures). Using primer pairs X-A, X-B, and X-C, fragments of expected size were obtained from the cDNA template utilized as a control and only a larger than cDNA size from the bovine genomic DNA (Figure 1A). Although other bands are present with primer X-B in the genomic DNA lane, these are nonspecific as only one was found by Southern blotting to hybridize to the cDNA probe (Figure 1B). The sizes of fragments amplified from genomic DNA and from the cDNA template are presented in Figure 1C. From these sizes and the locations of primers on the cDNA, we could predict the presence of at least two intervening sequences between the start of the mature protein-coding region and the poly(A) signal site, the first totaling 132 bp between X and A and a second totaling 361 bp between A and B (Figure 1C). This supposition was confirmed after subcloning and partial sequencing of the largest amplified product of 700 bp. This fragment was then labeled and utilized to screen a bovine genomic library in λEMBL-3. Three positive signals were identified. Southern blot analysis of these clones revealed that they were all identical. One of these, λ*COX7aH1*, was taken for further analysis.

The absence of cDNA-sized bands in the bovine genomic lanes suggests that there are no processed pseudogenes for *COX7aH*. This conclusion was also suggested by the simple genomic Southern blot pattern obtained when *COX7aH* cDNA was used as a probe (Seelan & Grossman, 1991a). By contrast, the liver isoform gene (*COX7aL*) is multicopy, most of them processed pseudogenes, as judged both by the complexity of the genomic Southern blot pattern (Seelan & Grossman, 1991a) and by the finding of cDNA-sized bands by PCR analysis of genomic DNA (Seelan and Grossman, unpublished observations).

Characterization of λ*COX7aH1*. An *Eco*RI digest of this clone revealed an unique 3.0-kb fragment (Figure 2A) that hybridized with the cDNA probe. Preliminary mapping and partial sequencing indicated that only about 1.86 kb of this fragment from the 5′ *Eco*RI end contained the complete gene sequence. To obtain additional sequences toward the 5′ region, an overlapping 4.7-kb *Xho*I frament (Figure 2A) was isolated that provided 3.18 kb of sequence upstream of the 5′ *Eco*RI site. This 3.18-kb *Xho*I–*Eco*RI fragment was cloned into pKS(−) and sequenced via subclones in its entirety. The restriction map and sequencing strategy of the *COX7aH* gene

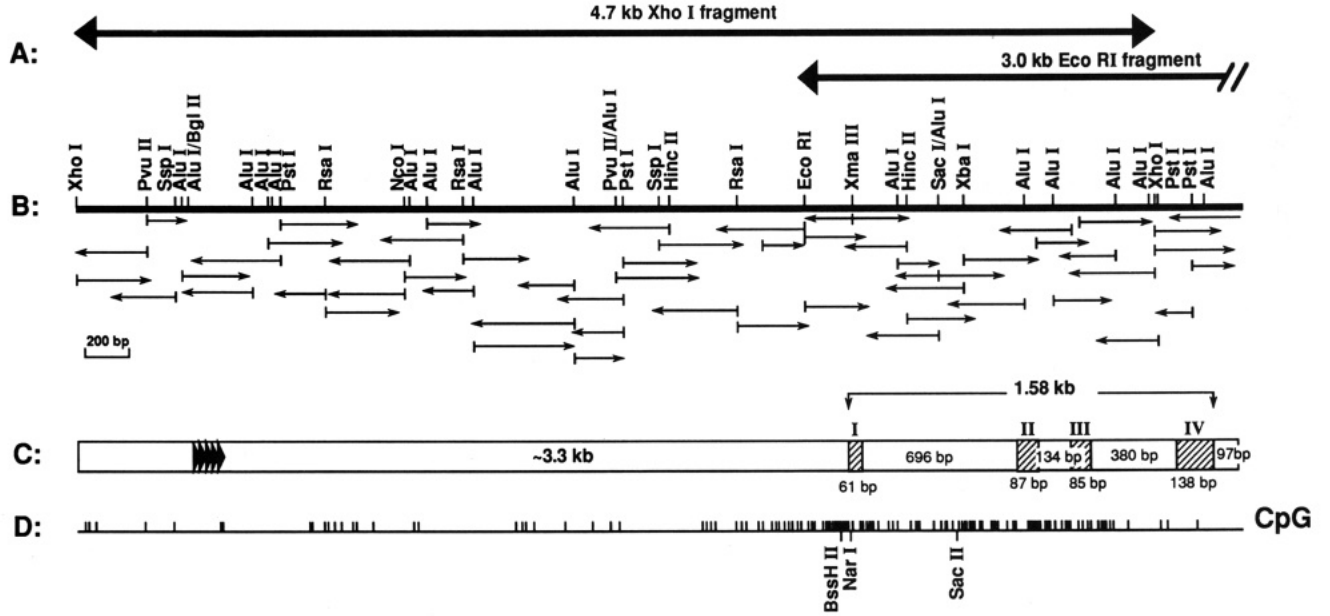


FIGURE 2: Characterization of the *COX7aH* gene and its upstream sequences. (A) Two overlapping genomic fragments—a 4.7-kb *Xho*I fragment and part of a 3-kb *Eco*RI fragment—were used for characterization. (B) Restriction map and sequencing strategy of the *COX7aH* gene and its upstream sequences. (C) The intron–exon organization of the gene. The four exons are hatched. Five 22-bp tandem direct repeats are located in the far upstream region (arrowheads). (D) Distribution of CpG residues. Three enzyme sites typically associated with CpG islands, *Bss*HII, *Nar*I, and *Sac*II, are present.

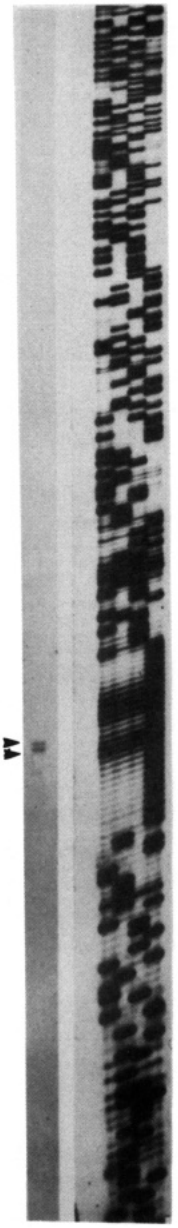


FIGURE 3: Primer extension analysis to determine the transcription start site. A primer complementary to the first 22 nt of the mature protein coding region was used in a reverse transcription reaction containing 50 μg of total bovine heart RNA. Primer-extended products were analyzed on 4% polyacrylamide sequencing gels. Two major products, 131 and 132 nt, were obtained (arrows). A sequencing reaction (left to right, A, C, G, T) run in parallel was used as a size marker. The presence of 29 T residues in the marker lane, which have the appearance of a compression, was confirmed by sequencing the complementary strand.

and its upstream regions are presented in Figure 2B.

Primary Structure of the Heart Gene. The 5.0-kb fragment illustrated in Figure 2C encompasses a region beginning about 3.3 kb upstream of the first exon to 97 bp downstream of the polyadenylation site. Exon I, which is 61 bp, contains the 5′ untranslated sequence and 5 amino acids of the presequence; exon II, 87 bp, encodes the remaining 16 amino acids of the presequence and the first 13 amino acids of the mature protein; exon III is 85 bp and encodes amino acids 14–41 of the mature protein; the 138-bp exon IV encodes the C-terminal 18 amino acids and contains the 3′ untranslated sequence. The four short exons are interspersed with three larger introns of sizes 696, 134, and 380 bp (Figure 2C). The location and size of the last two introns are consistent with what we inferred from our PCR analysis (Figure 1C). The splice-site junctions of all three introns follow the GT-AG rule (Mount, 1982).

Table I

ELEMENT	SEQUENCE[a]	MATCH	NT POS	STRAND
Motifs implicated in muscle regulation:				
1) CArG Box[b] (one GC bp allowed within the AT-rich center)	**CC (AT)$_6$ GG**			
	CC(TCAATT)GG	10/10	-1890	(+)
	CC(GTTTAT)GG	10/10	-1088	(+)
2) Muscle-specific consensus for binding sites (includes the MyoD motif-CANNTG)	**C$_N$GGCACCTGCC$_N$C** G AA GG TT G			
	CAGcCAGGTGaCTC	12/14	-1634	(+)
	GGGACAGCTGggGG	12/14	-971	(+)
	CAGcCAtCTGTTAC	12/14	-390	(+)
Motifs implicated in oxidative phosphorylation:				
1) OXBOX	**GGCTCTAAAGAGG**			
	GGCcCTcAgGAGG	10/13	-1968	(-)
2) "enhancer"	**TAGAGACAAGGTTTCACCA**			
	TAGAGcCcAtGcTTCACaA	14/19	-2277	(-)
	TgGcGctAgGGTTTCACCA	14/19	+1144	(+)
3) Hap2/Hap3 response element	**TN$^{A}_{G}$TTGGT**			
	TGGTTGGT	8/8	-1708	(+)
Others:				
1) GC box (Sp1 binding site)	**G$_{T}$GGGCGG$_{AAT}^{GGC}$**			
	TGGGCGGGGa	9/10	-103	(-)
	GGGGCGGGtT	9/10	+1133	(+)
2) SV40 core enhancer	**GTGG$^{AAA}_{TTT}$G**			
	GTGGTAAG	8/8	-2319	(+)
	GTGGAATG	8/8	-547	(-)
Repeats:				
1) Five tandem 22bp direct repeats	TcTTccTTggaGCATGCAGgtc	13/22[c]	-2870	(+)
	TTTTAgTTgcaGCATGCAGAAt	17/22	-2840	(+)
	TTTTATTTATTGCATGCAGAAA	22/22	-2818	(+)
	TTTTATTTATTGCATGCAGAAA	22/22	-2796	(+)
	TTTTAgTTgTatCATGtAGgAa	15/22	-2774	(+)
2) Two 10 bp direct repeats	AGAGGGGAGG	10/10	-23	(+)
	AGAGGGGAGG	10/10	+9	(+)

[a] Consensus sequences are in boldface type; lower-case letters indicate bases that differ from the consensus. [b] This motif is also present in nonmuscle genes. [c] As compared to the two perfect repeats; (+) and (−) indicate coding or noncoding strands, respectively.

The start of the first exon was determined by a primer extension assay using total bovine heart RNA and a primer specific to the first 22 bases of the mature heart protein. Two major products, 131 and 132 bases, were seen (Figure 3). The presence of two products differing in size by a single nt in a primer extension experiment is not unusual and has been attributed to a contaminating or inherent activity of the reverse transcriptase to add a nucleotide to the 5′ end in a template-independent fashion (Shelness & Williams, 1985). The 131-nt product was used to map the transcription start site (+1). Residues upstream of this site are negatively numbered.

The sequence of the *COX7aH* gene from −600 to +1678 is presented in Figure 4. Potential motifs typically found in genes involved in respiratory function and muscle gene activation are present both in the 5′-flanking region and in introns. The sequences and locations of these putative regulatory elements and their locations are presented in Table I and considered in the Discussion.

An analysis of the distribution of CpG residues (Figure 2D) over the entire 5.0-kb stretch reveals that this gene is associated with a CpG island (Bird, 1987). The 5.0-kb fragment contains a total of 120 CpGs, of which about 85 are located in the vicinity of the gene, extending from −360 to +1000 (Figure 4). The frequency of CpG residues within this region is about 1 per 18 bp whereas it is 1 per 100 bp in the remaining regions. This frequency is especially high in the region surrounding the transcriptional start site (−150 to +40), where the frequency is about 1 per 10 bp. Sites for typical CpG enzymes such as *Nar*I and *BssHII* are found immediately upstream of the gene,

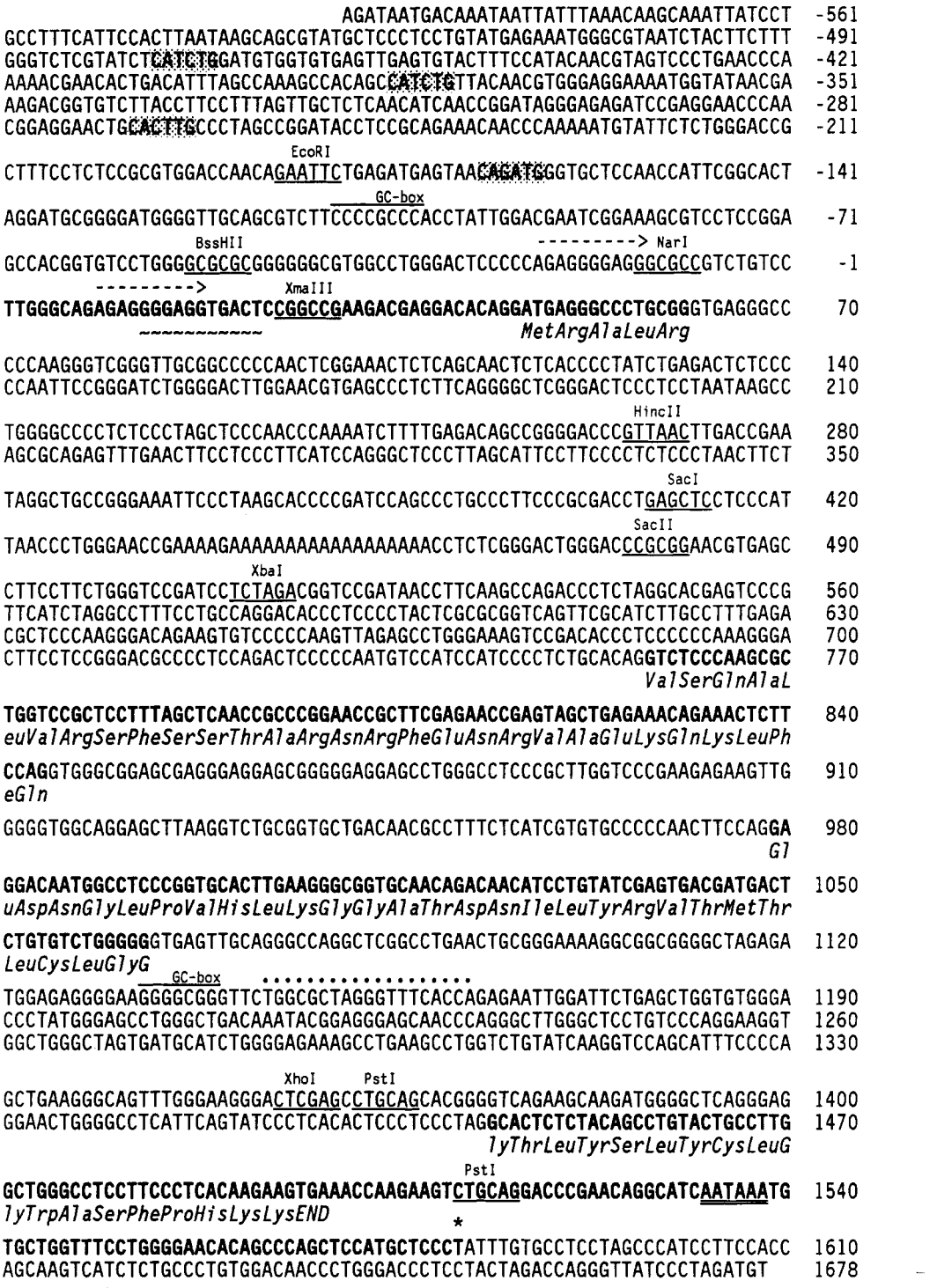


FIGURE 4: Nucleotide sequence of the *COX7aH* gene (−600 to +1678). The gene contains four exons (boldface) extending from the transcriptional start site (+1) to the polyadenylation site (*, at +1581). Highlighted in the sequence are restriction enzyme sites (underlined), GC boxes (overlined), potential MyoD motifs (shaded), the poly(A) signal (double underlined), the two 10-bp proximal repeats (arrows), and the respiratory enhancer (dotted lines). The wavy line represents the stretch of 11 nucleotides in the first exon not found in the cDNA (see text).

and a site for *Sac*II is found within the first intron (Figures 2D and 4).

Discussion

Three nuclear COX subunits—VIa, VIIa, and VIII—exist as H or L isoforms in most mammals examined. We have recently characterized the bovine cDNAs for each isoform of COX VIIa (Seelan et al., 1989; Seelan & Grossman, 1991a) and utilized them to show that transcription of *COX7aL* mRNA is observed in all tissues, whereas the mRNA for *COX7aH* is found exclusively in heart and skeletal muscle tissues. To understand the mechanisms of the differential regulation of these subunits and their cellular function, we have sought genomic clones encoding these isoforms. In this paper we present the isolation and characterization of the bovine *COX7aH* gene.

The gene consists of four exons that span 1.58 kb (Figure 2C). The 21 amino acid presequence is encoded in two exons—5 in exon I and 16 in exon II (Figure 5). It thus differs from the other presequence-containing COX subunit genes that have been analyzed so far—COX IV in rat (Yamada et al., 1990), mouse (Carter & Avadhani, 1991), and cow (Bachman et al., 1987) and COX Vb in mouse (Basu & Avadhani, 1991) and human (Lomax et al., 1991)—where the

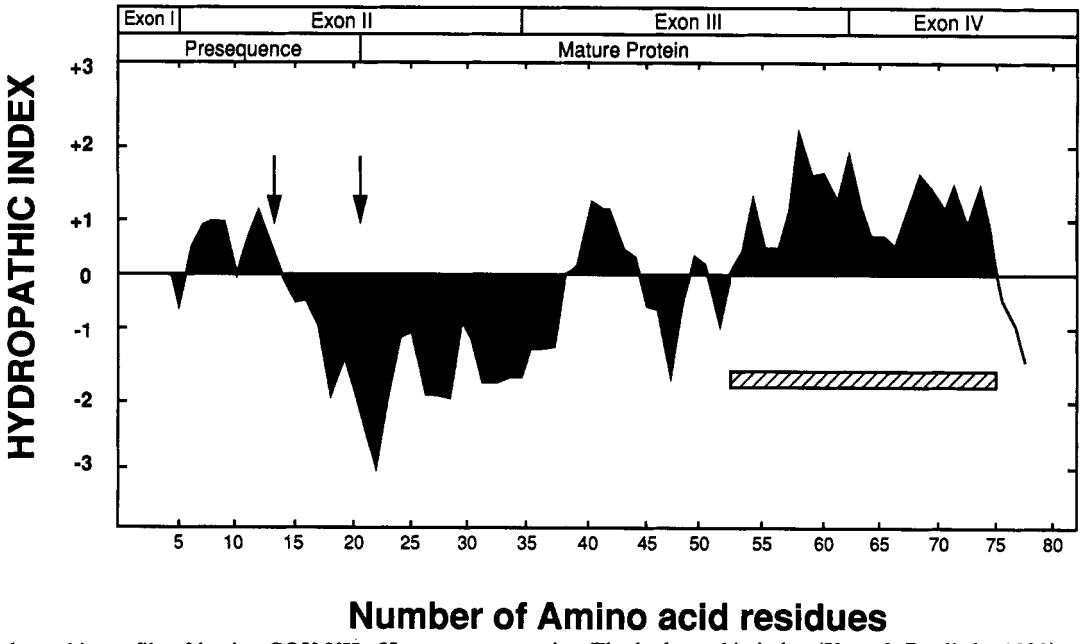


FIGURE 5: Hydropathic profile of bovine COX VIIa-H precursor protein. The hydropathic index (Kyte & Doolittle, 1982) was calculated with the running average obtained from a moving window of seven residues. The amino acid positions are indicated on the *X*-axis; the hydrophobic (+) or hydrophilic (−) indices are shown on the *Y*-axis. The locations of the four exons and the presequence region of the precursor protein are shown on top. The hatched area represents the putative membrane spanning domain, and the two arrows indicate presequence processing sites. The location of the first processing site (left) is based on a two-step processing model (Hendrick et al., 1989) inferred from the cDNA-derived sequence.

presequences are entirely contained within a single exon. Since mitochondrial presequences participate in a number of functions, such as binding to outer membrane receptors and directing the mature sequence to the correct intramitochondrial compartment, it would not be surprising if in at least some cases these multiple functions are represented by more than one domain. One possibility is that both domains are involved in presequence processing, since sites for a two-step processing mechanism of the precursor (Hendrick et al., 1989) are found in bovine COX VIIa-H (Seelan & Grossman, 1991a), with putative processing sites present after the 13th and 21st residue in the presequence (Figure 5). However, both processing sites are located in the middle of exon II, making it unlikely in this case that the split in the presequence is related to a processing domain.

The transmembrane region is also of interest in terms of exon boundaries. Although the N-terminal hydrophilic domain (Kyte & Doolittle, 1982) of the precursor protein is almost entirely contained within exon II, the putative transmembrane hydrophobic domain (Figure 5, hatched bar) appears to be split equally between exons III and IV. The transmembrane region is not a single distinct peak, however, but appears rather to consist of two overlapping hydrophobic peaks, with the trough between these peaks coinciding with the location of the third intron. It is possible that, during evolution, two domains encoding different functions have merged into the present-day protein. The presence of the hydrophobic sequence at the C-terminus, rather than at the N-terminus, also fits with the suggestion of Singer and Yaffe (1990) that such a feature would facilitate its import to the matrix, whereas a location closer to the presequence would favor its integration into the outer membrane.

Comparison of the cDNA and the genomic sequences reveals 11 bases in the untranslated region of the first exon (GGGAGGTGACT) (Figure 4, wavy line) that are not found in the bovine cDNA, where they are replaced with a 17-base sequence (AGGTAGTGGGGGAGAGC); sequences immediately upstream and downstream of this region are perfectly retained in both the cDNA and the genomic clones. However, the bovine genomic sequence in this region perfectly matches that found in the recently characterized human *COX7aH* cDNA (E. Arnaudo, M. Hirano, R. S. Seelan, A. Milatovich, C.-L. Hsieh, G. M. Fabrizi, L. I. Grossman, U. Francke, and E. Schon, submitted for publication); furthermore, on the basis of primer extension data, the bovine cDNA appears to extend beyond the deduced transcription start site by 4 nt. We believe that this sequence in the cDNA arose during cloning, presumably by acquiring additional bases either during cDNA construction or during PCR amplification. These results, however, do not affect the derived amino acid sequence of the heart precursor protein reported earlier (Seelan & Grossman, 1991a). The genomic sequence also confirms the presence of the three extra amino acids we observed at the C-terminus in the cDNA-derived sequence that were absent in the sequenced protein (Meinecke & Buse, 1986).

The *COX7aH* gene appears to contain several novel features. Although it is a tissue-specific gene, it contains no canonical TATA or CCAAT sequences upstream of the start site. AT-rich elements are found more than 300 bp upstream (Figure 4). Despite the primer extension results (Figure 3) that place the transcriptional start site at +1, we have not ruled out the existence of an intron in the immediate 5′ untranslated region.

The 5′ region of the gene is associated with a CpG island, a feature typical of many housekeeping genes. The nuclear gene encoding the human skeletal muscle/heart isoform of the mitochondrial adenine nucleotide translocator is another example of a respiratory gene that has CpG islands (Li et al., 1989; Cozens et al., 1989). CpG-island genes are generally presumed to be nonmethylated and hence transcriptionally active (Bird, 1986). For the tissue-specific human α-globin gene, it has been observed that the CpGs are nonmethylated regardless of whether the gene is expressed or not (Bird et al., 1987). Recent studies on chicken lysozyme (Wölfl et al., 1991) and the rat phosphoenolpyruvate carboxykinase (Faber et al., 1991) genes reveal a lack of correlation between cytosine

methylation in the promoter region and gene expression. We are currently examining the methylation status of this gene in various tissues.

The region immediately upstream, −1 to −100, is extremely G+C rich (72%) and contains a putative GC box (at −103; Gidoni et al., 1984), in the opposite orientation, having a 9/10 match with the consensus sequence. Another potential motif is located in the third intron at +1133 (Table I; Figure 4). GC boxes are potential sites for the binding of the transcription factor Sp1, which is known to enhance RNA Pol II transcription severalfold (Kadonaga et al., 1986). The upstream region also contains two sequences (at −547 and −2319) that show perfect homology with the SV40 core enhancer (Weiher et al., 1983) (Table I).

Since the *COX7aH* gene is expressed in muscle tissues, we have scanned the upstream sequences for possible muscle regulatory motifs. The myogenic program in mammalian skeletal muscle is initiated by MyoD and related myogenic factors (Weintraub et al., 1991). The simple sequence CANNTG, implicated in MyoD binding, has been observed in the regulatory regions of several muscle-specific genes (Lassar et al., 1989). At least four sites (Figure 4, shaded) can be localized between −1 and −600 of the *COX7aH* gene. One of these, at −390, and two others, at −971 and −1634 (Table I), match the 14-nt consensus sequence proposed for binding sites in muscle-specific regulatory regions (Buskin & Hauschka, 1989). A third element, a decameric sequence designated the CArG box, CC(AT)$_6$GG, which is a binding site for the ubiquitously expressed serum response factor (SRF), has been shown to play an important role in the tissue-specific expression of the human cardiac α-actin gene in skeletal muscle (Sartorelli et al., 1990). CArG boxes, where one allowable GC bp can occur within the AT-rich center, have been found in the upstream regions of genes encoding the chicken and human α-cardiac actin; rat and chicken α-skeletal actin; human, chicken, and rat β-actin; rat α-cardiac myosin heavy chain; chicken cardiac and rat skeletal myosin light chain 2; chicken cardiac troponin T (Minty & Kedes, 1986); and mouse muscle creatine kinase (Sternberg et al., 1988). Two perfect motifs at −1890, CC(TCAATT)GG, and −1088, CC(GTTTAT)GG, are found in the *COX7aH* flanking regions. Interactions involving the GC box, the CArG box, and the MyoD element have been implicated in the muscle-specific expression of the human cardiac actin gene (Sartorelli et al., 1990). In addition to the upstream sequences, functionally important muscle regulatory domains could also reside in the large first intron of *COX7aH*, as has been observed for several muscle genes with such a feature (Shen et al., 1991).

Although the above features may account for *COX7aH* expression in skeletal muscle, it may not explain its expression in cardiac tissues, where MyoD and related myogenic factors are not synthesized (Davis et al., 1987; Sassoon et al., 1989; Shen et al., 1991). Thus, the gene is likely to contain additional sequence elements that confer cardiac muscle expression. Such a situation is seen to exist for the cardiac troponin T gene, where a minimum promoter region is sufficient for expression in skeletal muscle, whereas an additional upstream sequence is required for expression in cardiac myocytes (Mar & Ordahl, 1990). Similarly for the human cardiac actin gene, multiple CArG boxes in the upstream region have been implicated to play a major role in cardiac muscle expression, whereas a core promoter located downstream comprising a CArG box, Sp1, and MyoD elements plays a crucial role in skeletal muscle expression (Sartorelli et al., 1990). On the basis of the foregoing observations, the two CArG boxes located far upstream (−1088 and −1890) in *COX7aH* could mediate cardiac muscle expression, whereas the Sp1 binding site (−103) and the 14-nt consensus MyoD site (−391), in association with the CArG box(es), could confer skeletal muscle expression.

An alternative mechanism by which a gene could be expressed both in skeletal and in cardiac muscle is that a cardiac-specific factor could recognize and bind to the MyoD regulatory element. Analysis of the human cardiac α-actin promotor reveals that the sequences essential for expression in cardiac myocytes are similar to those implicated in MyoD-induced enhancement, suggesting that cardiac muscle might contain a MyoD-like activity (Pari et al., 1991). That two different factors could bind to the same consensus sequence has been observed for the basic helix–loop–helix (bHLH) family of transcriptional proteins (Weintraub et al., 1991).

In addition to muscle-specific elements, transcriptional control elements specific for genes of the respiratory complex have recently been identified [Li et al., 1990; Virbasius & Scarpulla, 1991; reviewed by Nagley (1991)]. These include (i) the OXBOX, a 13-nt tissue-specific regulatory element that is found to enhance transcription in at least two respiratory genes expressed maximally in heart and skeletal muscle (Li et al., 1990); (ii) a NRF-1 (nuclear respiratory factor 1) site that is believed to coordinate the regulation of nuclear and mitochondrial genomes (Evans & Scarpulla, 1989, 1990); (iii) NRF-2 sites that have been implicated in the expression of rat COX IV and human ATP synthase β-subunit genes (Virbasius & Scarpulla, 1991); (iv) an "enhancer" element specific for respiratory proteins (Tomura et al., 1990); and (v) a set of cis elements termed Mt1, Mt2, Mt3, and Mt4 (Suzuki et al., 1989, 1990, 1991); the latter two also occur in the D-loop promoter region of mitochondrial DNA. The Mt1, Mt3, and Mt4 elements show positional correlation within a region of 250–300 bp in the upstream regions of at least four respiratory genes (Suzuki et al., 1990).

We found no matches with the NRF-1 (allowing for two mismatches) or NRF-2 (allowing for a single mismatch) consensus sequences, but a sequence in the reverse orientation bearing a 10/13 relationship with the OXBOX is present at −1968 (Table I). Two motifs that resemble the enhancer element (14/19 match) are found upstream, at −2277, and in the third intron, at +1144 (Figure 4; Table I); the latter has a stretch of 10 identical nucleotides at its 3′ end and is located adjacent to a potential GC box. Although a few putative Mt2, Mt3, and Mt4 elements are present in the upstream region (not shown), there is only one Mt1-like sequence (TATTCAGGg at −999, in the reverse orientation). However, none of the MT3 or MT4 sequences show any positional correlation with this element. The *COX7aH* gene also contains the octanucleotide TGGTTGGT at −1708, which perfectly matches the consensus (TNRTTGGT) for the yeast Hap 2/Hap 3 responsive element (Olesen et al., 1987; Maarse et al., 1988). This sequence has been found in the 5′-flanking regions of many yeast genes involved in respiratory function and in a few animal genes encoding heme-related proteins; included among the latter are the rat and mouse COX IV, mouse COX Vb, rat COX VIc, and human cytochrome c_1 genes [cited in Yamada et al. (1990) and Carter and Avadhani (1991)].

In addition, there are two sets of repeat elements (Table I). The first comprises five 22-bp tandem direct, including two perfect, repeats located ~2.7 kb upstream, and the second consists of a pair of 10-bp direct repeats spaced 21 bp apart in the vicinity of the transcription start site (−23 to −14 and +9 to +18). Interestingly, two 11-bp repeats have also been identified in the rat COX IV gene between the minor and

major transcriptional start sites (Virbasius & Scarpulla, 1991); these repeats are the binding sites for a protein, designated as NRF-2. An NRF-2 site functional in a DNA binding assay is also found in the first intron of the ATP synthase β-subunit gene (Virbasius & Scarpulla, 1991). The 10-bp repeats in *COX7aH*, however, show no sequence similarity to the NRF-2 binding sites. It remains to be seen whether these sites have any role in respiratory function.

The function of bovine COX VIIa-H, like those of the other nuclear-encoded subunits, remains to be elucidated. There is indirect evidence to suggest that it may play a crucial role in COX function—the presence of this subunit (and COX VIIb) was not detectable in muscle tissues in fatal and benign mitochondrial myopathies (Tritschler et al., 1991). In addition, COX VIIa-H appears to be developmentally regulated (unpublished data). Analysis of the bovine gene reveals several elements characteristic of muscle, housekeeping, and respiratory functions. Although the functionality of these elements remains to be tested, the sequence data suggest that this gene is under complex control. This is to be expected for a gene that is transcribed only in skeletal and cardiac muscle tissues and whose expression has to be coordinately regulated with those of the other nuclear and mitochondrial COX subunits. The characterization of this gene and its upstream sequences now paves the way for an elucidation of its mode of regulation, and perhaps function, in mammalian cells.

ACKNOWLEDGMENTS

We thank Dr. Margaret I. Lomax (University of Michigan) for making available filters containing a λEMBL-3 genomic library.

REFERENCES

Anthony, G., Stroh, A., Lottspeich, F., & Kadenbach, B. (1990) *FEBS Lett. 277*, 97–100.

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., & Struhl, K., Eds. (1989) *Current Protocols in Molecular Biology*, John Wiley and Sons, New York.

Bachman, N. J., Lomax, M. I., & Grossman, L. I. (1987) *Gene 55*, 219–229.

Basu, A., & Avadhani, N. (1991) *J. Biol. Chem. 266*, 15450–15456.

Bird, A. P. (1986) *Nature 321*, 209–213.

Bird, A. P. (1987) *Trends Genet. 3*, 342–347.

Bird, A. P., Taggart, M. H., Nicholls, R. D., & Higgs, D. R. (1987) *EMBO J. 6*, 999–1004.

Bisson, R., Schiavo, G., & Papini, E. (1985) *Biochemistry 24*, 7845–7852.

Buskin, J. N., & Hauschka, S. D. (1989) *Mol. Cell. Biol. 9*, 2627–2640.

Capaldi, R. A. (1990) *Arch. Biochem. Biophys. 280*, 252–262.

Capaldi, R. A., Takamiya, S., Zhang, Y.-Z., Gonzalez-Halphen, D., & Yanamura, W. (1987) *Curr. Top. Bioenerg. 15*, 91–112.

Carter, R. S., & Avadhani, N. G. (1991) *Arch. Biochem. Biophys. 288*, 97–106.

Ciccarelli, R. B., & Wetterhahn, K. E. (1984) *Cancer Res. 44*, 3892–3897.

Cozens, A. L., Runswick, M. J., & Walker, J. E. (1989) *J. Mol. Biol. 206*, 261–280.

Davis, R. L., Weintraub, H., & Lassar, A. B. (1987) *Cell 51*, 987–1000.

Dowhan, W., Bibus, C. R., & Schatz, G. (1985) *EMBO J. 4*, 179–184.

Evans, M. J., & Scarpulla, R. C. (1989) *J. Biol. Chem. 264*, 14361–14368.

Evans, M. J., & Scarpulla, R. C. (1990) *Genes Dev. 4*, 1023–1034.

Faber, S., Ip, P., Granner, D., & Chalkley, R. (1991) *Nucleic Acids Res. 19*, 4681–4688.

Feinberg, A. P., & Vogelstein, B. (1983) *Anal. Biochem. 132*, 6–13.

Gidoni, D., Dynan, W. S., & Tjian, R. (1984) *Nature 312*, 409–413.

Hendrick, J. P., Hodges, P. E., & Rosenberg, L. E. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 4056–4060.

Kadenbach, B. (1986) *J. Bioenerg. Biomembr. 18*, 39–54.

Kadenbach, B., & Merle, P. (1981) *FEBS Lett. 135*, 1–11.

Kadenbach, B., Jarausch, J., Hartmann, R., & Merle, P. (1983) *Anal. Biochem. 129*, 517–521.

Kadenbach, B., Kuhn-Nentwig, L., & Büge, U. (1987) *Curr. Top. Bioenerg. 15*, 113–161.

Kadenbach, B., Reimann, A., Stroh, A., & Hüther, F.-J. (1988) in *Oxidases and Related Redox Systems*, pp 653–668, Alan R. Liss, Inc., New York.

Kadenbach, B., Schlerf, A., Mengel, T., Hengst, L., Cao, X., Suske, G., Eckerskorn, C., & Lottspeich, F. (1989) in *Organelles in Eukaryotic Cells: Molecular Structure and Interactions* (Tager, J. M., Azzi, A., Papa, S., & Guerrieri, F., Eds.) pp 143–156, Plenum, New York and London.

Kadonaga, J. T., Jones, K. A., & Tjian, R. (1986) *Trends Biochem. Sci. 11*, 20–23.

Kennaway, N. G., Carrero-Valenzuela, R. D., Ewart, G., Balan, V. K., Lightowlers, R., Zhang, Y.-Z., Powell, B. R., Capaldi, R. A., & Buist, N. R. M. (1990) *Pediatr. Res. 28*, 529–535.

Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol. 157*, 105–132.

Lassar, A. B., Buskin, J. N., Lockshon, D., Davis, R. L., Apone, S., Hauschka, S. D., & Weintraub, H. (1989) *Cell 58*, 823–831.

Li, K., Warner, C. K., Hodge, J. A., Minoshima, S., Kudoh, J., Fukuyama, R., Maekawa, M., Shimizu, Y., Shimizu, N., & Wallace, D. C. (1989) *J. Biol. Chem. 264*, 13998–14004.

Li, K., Hodge, J. A., & Wallace, D. C. (1990) *J. Biol. Chem. 265*, 20585–20588.

Lightowlers, R., Ewart, G., Aggeler, R., Zhang, Y.-Z., Calavetta, L., & Capaldi, R. A. (1990) *J. Biol. Chem. 265*, 2677–2681.

Lomax, M. I., & Grossman, L. I. (1989) *Trends Biochem. Sci. 14*, 501–503.

Lomax, M. I., Hsieh, C.-L., Darras, B. T., & Francke, U. (1991) *Genomics 10*, 1–9.

Maarse, A. C., de Haan, M., Bout, A., & Grivell, L. A. (1988) *Nucleic Acids Res. 16*, 5797–5811.

Malatesta, F., Antonini, G., Sarti, P., & Brunori, M. (1987) *Biochem. J. 248*, 161–165.

Mar, J. H., & Ordahl, C. T. (1990) *Mol. Cell. Biol. 10*, 4271–4283.

Meinecke, L., & Buse, G. (1986) *Biol. Chem. Hoppe-Seyler 367*, 67–73.

Merle, P., & Kadenbach, B. (1982) *Eur. J. Biochem. 125*, 239–244.

Mierendorf, R. C., & Pfeffer, D. (1987) *Methods Enzymol. 152*, 559–560.

Minty, A., & Kedes, L. (1986) *Mol. Cell. Biol. 6*, 2125–2136.

Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459–472.

Nagley, P. (1991) *Trends Genet. 7*, 1–4.

Olesen, J., Hahn, S., & Guarente, L. (1987) *Cell 51*, 953–961.

Pari, G., Jardine, K., & McBurney, M. W. (1991) *Mol. Cell. Biol. 11*, 4796–4803.
Power, S. D., Lochrie, M. A., Sevarino, K. A., Patterson, T. E., & Poyton, R. O. (1984) *J. Biol. Chem. 259*, 6564–6570.
Rizzuto, R., Nakase, H., Darras, B., Francke, U., Fabrizi, G. M., Mengel, T., Walsh, F., Kadenbach, B., DiMauro, S., & Schon, E. A. (1989) *J. Biol. Chem. 264*, 10595–10600.
Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., & Erlich, H. A. (1988) *Science 239*, 487–491.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.
Sartorelli, V., Webster, K. A., & Kedes, L. (1990) *Genes Dev. 4*, 1811–1822.
Sassoon, D., Lyons, G., Wright, W. E., Lin, V., Lassar, A., Weintraub, H., & Buckingham, M. (1989) *Nature 341*, 303–307.
Schlerf, A., Droste, M., Winter, M., & Kadenbach, B. (1988) *EMBO J. 7*, 2387–2391.
Seelan, R. S., & Grossman, L. I. (1991a) *J. Biol. Chem. 266*, 19752–19757.
Seelan, R. S., & Grossman, L. I. (1991b) *BioTechniques 10*, 186–188.
Seelan, R. S., Scheuner, D., Lomax, M. I., & Grossman, L. I. (1989) *Nucleic Acids Res. 17*, 6410.
Shelness, G. S., & Williams, D. L. (1985) *J. Biol. Chem. 260*, 8637–8646.
Shen, R., Goswami, S. K., Mascareno, E., Kumar, A., & Siddiqui, M. A. Q. (1991) *Mol. Cell. Biol. 11*, 1676–1685.
Singer, S. J., & Yaffe, M. P. (1990) *Trends Biochem. Sci. 15*, 369–373.
Sternberg, E. A., Spizz, G., Perry, W. M., Vizard, D., Weil, T., & Olson, E. N. (1988) *Mol. Cell. Biol. 8*, 2896–2909.
Suske, G., Enders, C., Schlerf, A., & Kadenbach, B. (1988) *DNA 7*, 163–171.
Suzuki, H., Hosokawa, Y., Nishikimi, M., & Ozawa, T. (1989) *J. Biol. Chem. 264*, 1368–1374.
Suzuki, H., Hosokowa, Y., Toda, H., Nishikimi, M., & Ozawa, T. (1990) *J. Biol. Chem. 265*, 8159–8163.
Suzuki, H., Hosokowa, Y., Nishikimi, M., & Ozawa, T. (1991) *J. Biol. Chem. 266*, 2333–2338.
Taanman, J.-W., Schrage, C., Bokma, E., Reuvekamp, P., Agsteribbe, E., & De Vries, H. (1991) *Biochim. Biophys. Acta. 1089*, 283–285.
Takamiya, S., Lindorfer, M. A., & Capaldi, R. A. (1987) *FEBS Lett. 218*, 277–282.
Tomura, H., Endo, H., Kagawa, Y., & Ohta, S. (1990) *J. Biol. Chem. 265*, 6525–6527.
Tritschler, H. J., Bonilla, E., Lombes, A., Andreetta, F., Servidei, S., Schneyder, B., Miranda, A. F., Schon, E. A., Kadenbach, B., & DiMauro, S. (1991) *Neurology 41*, 300–305.
Van Beeumen, J. J., Van Kuilenburg, A. B. P., Van Bun, S., Van den Bogert, C., Tager, J. M., & Muijsers, A. O. (1990) *FEBS Lett. 263*, 213–216.
Van Kuilenburg, A. B. P., Muijsers, A. O., Demol, H., Dekker, H. L., & Van Beeumen, J. J. (1988) *FEBS Lett. 240*, 127–132.
Virbasius, J. V., & Scarpulla, R. C. (1990) *Nucleic Acids Res. 18*, 6581–6586.
Virbasius, J. V., & Scarpulla, R. C. (1991) *Mol. Cell. Biol. 11*, 5631–5638.
Weiher, H., Konig, M., & Gruss, P. (1983) *Science 219*, 626–631.
Weintraub, H., Davis, R., Tapscott, S., Thayer, M., Krause, M., Benezra, R., Blackwell, T. K., Turner, D., Rupp, R., Hollenberg, S., Zhuang, Y., & Lassar, A. (1991) *Science 251*, 761–766.
Wölfl, S., Schrader, M., & Wittig, B. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 271–275.
Yamada, M., Amuro, N., Goto, Y., & Okazaki, T. (1990) *J. Biol. Chem. 265*, 7687–7692.