Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression

Laurent Duret, Franck Dorkeld and Christian Gautier
Laboratoire de Biométrie, Génétique et Biologie des Populations, Université Claude Bernard, Lyon I,
URA-CNRS 243 Bat 741, 43 Bd du 11 Novembre 1918, 69622 Villeurbanne cedex, France

Received March 11, 1993; Accepted April 16, 1993

ABSTRACT

Comparison of nucleotide sequences from different classes of vertebrates that diverged more than 300 million years ago, revealed the existence of highly conserved regions (HCRs) with more than 70% similarity over 100 to 1450 nt in non-coding parts of genes. Such a conservation is unexpected because it is much longer and stronger than what is necessary for specifying the binding of a regulatory protein. HCRs are relatively frequent, particularly in genes that are essential to cell life. In multigene families, conserved regions are specific of each isotype and are probably involved in the control of their specific pattern of expression. Studying HCRs distribution within genes showed that functional constraints are generally much stronger in 3′-non-coding regions than in promoters or introns. The 3′-HCRs are particularly A + T-rich and are always located in the transcribed untranslated regions of genes, which suggests that they are involved in post-transcriptional processes. However, current knowledge of mechanisms that regulate mRNA export, localisation, translation, or degradation is not sufficient to explain the strong functional constraints that we have characterised.

INTRODUCTION

While constraints that affect the rate of accumulation of non-synonymous mutations in the coding part of genes are quite well understood, little is known about those affecting non-coding regions, that are often considered as neutral. The non-coding regions (NCRs) of a gene include the 5′ flanking (untranscribed) region (5′-FLR), 5′ untranslated region (5′-UTR), 3′ untranslated region (3′-UTR), 3′-flanking region (3′-FLR), and introns. Interestingly, numerous authors have described genes with very high conservation in non-coding sequences between distantly related species. Yaffe *et al.* (1) were the first to discuss this phenomenon; they have shown that portions of the 3′-UTRs of skeletal muscle alpha-actin and cytoplasmic beta-actin had been preserved since the divergence of mammals and birds and that the conservation was specific of each actin isotypes. Striking examples of conservation between mammals and birds include the dystrophin gene 3′-UTR, that contains three regions of respectively 473, 113 and 343 nt with more than 80% similarity (2), the histone replacement variant H3.3 gene entire 3′-UTR (520 nt) with 85% similarity (3) or the insulin-like growth factor I 5′-NCR (80% over 450 nt) (4). We recently reported the existence of an exceptionally long conserved region (82% similarity over 930 nt) in the 3′-UTR of the human and chicken BTG1 antiproliferative genes (5). Such a conservation implies a strong selective pressure. However, long regions with high sequence similarity are generally not associated with binding sites of regulatory proteins, which generally require less specificity over a smaller region. Various steps of gene expression involve the binding of factors on short signals (TATA-box, polyadenylation signal etc...) or the formation of particular RNA spatial structure such as the stem loop structure at the end of replication dependent histones mRNAs (6). However these mechanisms involve sequences that are much shorter than the conserved NCRs mentioned above. A study of 11 mammalian genes has shown that introns, 3′-FLR and fourfold degenerate sites evolve almost at the same rate as pseudogenes (0,5% per million years) that are thought to be free of any selective pressure, 2.5 times faster than 5′-FLR, 5′-UTR and 3′-UTR and 5 times faster than non-degenerate sites (7). This variability of evolutionary rates between each part of genes reflects the different functional constraints that operate on it (7). However, the isolated examples of high conservation that have been reported, do not allow to draw conclusion about their function.

Since the estimated rate of accumulation of neutral point substitution during evolution is about 0,5% per million years (My) (7), the sequence similarity between species that diverged 300 My ago in DNA regions which are not subject to selective pressure should be about 30% (after correction for multiple substitutions), approximately the same as between unrelated sequences. The evolutionary distances between mammals and birds, amphibians or fishes are respectively 300 My, 360 My and more than 420 My (8,9). Therefore, these distances are well appropriated to detect regions that are subject to a strong selective pressure. Thanks to the growing amount of published vertebrate sequences available through databases, it is now possible to systematically search for highly conserved regions and study the constraints that affect non-coding sequences.

With this objective, we reviewed all the genes included in GenBank database which contain long portions of their non-coding regions that remain highly conserved between species belonging to different classes of vertebrates. We show that such regions are not exceptional, and appear preferentially in genes that are essential to cell life. The study of their distribution within genes reveals that most of them are involved in post-transcriptional mechanisms. We discuss the potential role of these sequences in the control of mRNA export, localisation, translation, or destabilization.

MATERIALS AND METHODS

Search for homologous sequences in non-coding regions

All sequences of at least 100 nt situated downstream of the stop codon of vertebrate protein-coding genes were extracted from GenBank database (10, release 72, June 15 1992) using ACNUC retrieval software (11). With this subset, we created three databases: the first containing the mammalian sequences, the second the avian sequences and the third the sequences from the other classes of vertebrates (namely: amphibian, reptile, osteochthyes, chondrichthyes and agnatha). We compared the three databases with each other with blastn (basic local alignment search tool—version for nucleic-acids sequences) (12). The cut-off score (S) for reporting pairs of homologous sequences was set to 150, with [$S = 5\times A-4\times B$] where A is the number of identical nucleotides between the two sequences and B is the number of mismatches. Homologous 3′ non-coding region were then aligned with CLUSTALV (13). The similarity was calculated as 100 × the ratio between matching nucleotides and total number of nucleotides in the sequence. Each gap was scored as one mismatch and therefore was counted as one nucleotide in the overall calculation of the length of sequence. In order to delimit the regions of at least 100 nt with 70% similarity or more, we drew the profile of homology between two sequences by computing the similarity on a 100-nt window moving along the alignment at 10-nt intervals. The same procedure was applied to all 5′ non-coding regions and introns of at least 100 nt.

Search for orthologous protein genes

All complete vertebrate coding sequences from GenBank were translated, and the couples of homologous protein genes from different classes of vertebrates were searched with blastp (12). Then couples of homologous but not orthologous genes were eliminated on the basis of information included in GenBank. In an attempt to conserve only orthologous genes, those which could not be clearly identified among different members of a multigene family were also discarded.

Sequence analysis

Sequence analysis (G+C content, search for ORF etc.....) were performed with Analseq (14). The isochore in which a gene is located was predicted according to its G+C-content in codons third position (G+C%III) as previously described (15): G+C%III < 57% for L1+L2 isochores, G+C%III < 75% for H1+H2 isochores, G+C%III ≥75% for H3 isochore.

RESULTS AND DISCUSSION

Highly conserved sequences in non-coding regions between different classes of vertebrates

Sequence comparisons of evolutionarily related genes may reveal different patterns of biologically significant conservation: a region of strong similarity over a short sequence (<15 nt) may correspond to the target of regulatory DNA- or RNA-binding proteins, such as signals involved in polyadenylation, splicing, transcription or translation; conversely, weak similarity over a long region (e.g. >100 nt) may reveal sequences coding for distantly related proteins. In this work we focused our attention upon regions with a high degree of conservation, that is to say that are subjected to strong selective pressure, and that are much longer than what is generally necessary to specify the binding of a protein. Therefore we defined a highly conserved region (HCR) as a sequence of at least 100 nt with 70% similarity or more between species that diverged more than 300 My ago (i.e. that should share only 30% similarity in absence of selective pressure). 5′- and 3′-HCRs will refer to HCRs found respectively in 5′- or 3′-non-coding regions.

We compared all the 3′-, 5′-non-coding sequences and introns of at least 100 nt from different classes of vertebrates. In order to be exhaustive, we chose a low cut-off parameter (S=150, see material and methods) so that we could detect homologous regions with about 60% similarity over 100 nt. We eliminated couples of sequences for which similarity was due to microsatellites. Microsatellites consist in repeats of short oligonucleotides (generally 2–4 nt) that are widely distributed among mammalian genomes (16) and, as we observed, are present as well in genomes from other vertebrates classes. We also eliminated couples for which we found homology between long terminal repeats of integrated retroviruses or artificial cloning vectors included by error in the published sequence. Finally, we eliminated couples for which the homologous region corresponds to coding exons that lie 3′ of the stop codon as a result of alternative splicing, presence of an adjacent gene or even errors in sequence annotation in the database. Remaining homologous regions correspond to unique sequences, specific of one gene or gene family (see below), and whose position in the gene has been conserved during evolution. Therefore it was possible to align homologous non-coding regions. In order to study the pattern of conservation, we calculated the profile of similarity for each alignment, and we delimited regions with more than 60%, 70% or 80% similarity over 100 nt. Table 1 gives the list of the 87 genes for which we detected HCRs. In almost all cases, the conserved region is specific of a single gene. The only exceptions are the short 5′-HCR found in both murine Hox1.1 and Hox2.3 genes [74, 86] and the 3′-HCR in N- and S-myc [5] (numbers in brackets refer to the numbering used in table 1). Examples of striking patterns of similarity are presented in figure 1. The cumulative length of HCRs within a gene ranges from 100 to 1450 nt (in the 3′-UTR of the endoplasmic-reticulum Ca(2+)-transport ATPase SERCA2B of pig and chicken) (fig. 1a, [1]). The longest continuous HCR was found between human and chicken BTG1 antiproliferative gene 3′-UTR, with 82% similarity over 930 nt (fig. 1a, [3]).

Conserved regions do not correspond to coding sequences

The simplest explanation for these conserved regions would be that they represent protein-encoding exons from adjacent or overlapping genes, or involved in the coding of another protein isotype via alternative splicing.

To assess this possibility, entire non-coding regions of all HCR-containing genes were first compared with all protein-encoding sequences described in GenBank. Apart from exceptions mentioned above, we did not find any significant homology. Then, we translated all open reading frames (ORFs) of at least 75 nt from non-coding sequences in the six possible frames, and these potential polypeptides were compared to each other. By this way, we found 80 potential coding sequences preserved

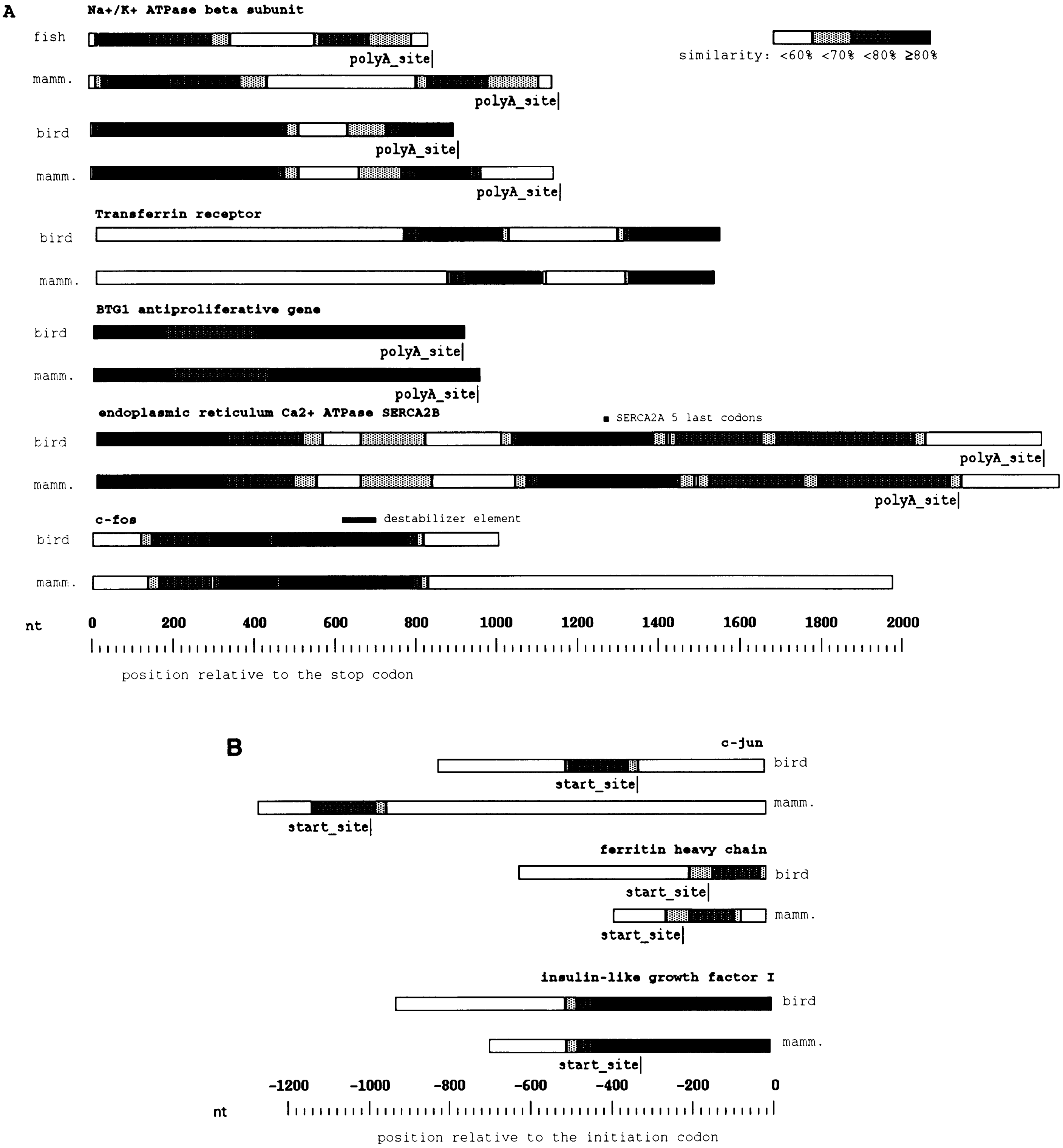

Figure 1. Pattern of similarity between non-coding regions from homologous genes belonging to different vertebrate classes. **A.** 3′ non-coding regions; **B.** 5′ non-coding regions. The position of polyadenylation sites or transcription start sites are indicated when known. The GenBank accession numbers of sequences are listed in table 1.

during vertebrates evolution. However, in most cases these conserved ORFs concern only a small portion of the HCR, the longest coding only for a 46 amino-acids protein. We further analysed the 20 longest conserved ORFs, and we found that either there was much more similarity between the sequences at the DNA level than at the potential protein level due to the presence of frame-shift mutations or that the similarity was higher or equal in the third codon positions compared with the first or second. This situation is inverse to that usually observed for related protein-encoding DNAs, where mutations accumulate mainly at silent sites and where gaps are generally observed by multiple of three nt for maintaining the reading frame. Thus it is very

Table 1. High similarity between non-coding regions from homologous genes belonging to different classes of vertebrates

GENE	GenBank Accession Number				Non-coding region common sequenced length (nt) (a). mammals versus:					
	mamm.	birds	amph.	fishes	birds		amph.		fishes	
					HCR	total	HCR	total	HCR	total
3'-HCR-containing genes										
1 endoplasmic-reticulum Ca2+ ATPase SERCA2B	X53754	M66385			1417	2364				
2 dystrophin	M18533	X13369			1024	2443				
3 BTG1 antiproliferative gene	X61123	X64146			923	923				
4 Neural Cell Adhesion Molecule (b)	X15051	M15861	M25696		769	3306	0	430		
5 N-myc (d)	X53674	D90071			741	1082				
6 smooth muscle caldesmon	M64110	J04968			667	1248				
7 c-fos	K00650	M37000			661	1026				
8 Na/K-ATPase beta subunit (b)	X03747	J02787		X03471	656	920			405	860
9 histone replacement variant H3.3B (f)	X13605	M11393			616	836				
10 calmodulin II (b,f)	M17069	L00101	K01945		551	593	155	279		
11 HMG-17	X13546	J03229			543	1259				
12 histone replacement variant H3.3A (f)	X05857	M11667			533	694				
13 MARCKS	M24638	M31650			532	632				
14 alpha-2 collagen gene type I	K01078	M25984			528	984				
15 middle weight neurofilament	Y00067	X17102			505	672				
16 transferrin receptor	M58040	X13753			450	1543				
17 c-jun	J04111	X15547			439	624				
18 GATA-3 transcription factor	X58072		M76565				435	796		
19 N-cadherin	X57548	X07277			431	431				
20 calbindin-D28K	X06661	X06633			419	1568				
21 Y-box binding protein 1	M57299		M59453				383	386		
22 lysosomal membrane glycoprotein	M32015	M59365			371	776				
23 ornithine decarboxylase (b)	X16277	X64710	X56316		333	334	0	281		
24 NFI-B/nuclear factor I-A1	D90173	X51486			316	479				
25 alpha-1 collagen type II	X06268	L00063	M63596		117	441	314	441		
26 vimentin (b)	M25246	V00447	X16843		310	323	0	323		
27 elongation factor 1-alpha (f)	J04617		M25697				289	332		
28 78-kD glucose-regulated protein	M19645	M27260			288	374				
29 cytoskeletal gamma-actin (f)	X52815	M26111			279	730				
30 peptidylhydroxyglycine N-C lyase	M25732		X62771				244	546		
31 contactin	X14943	Y00813			239	532				
32 transcription factor 4 (CTF4)	M83233	M87337			223	223				
33 HMG-14	M21339	X52708			220	2018				
34 platelet-derived growth factor A	X06374		M23238				218	802		
35 retinoic acid binding protein beta	Y00291	X57340			209	229				
36 cartilage link protein	Y00165	M13212			188	253				
37 cardiac alpha-actin (b,f)	J00073	M10607	X04669		187	374	(c)	374		
38 c-myb	M15024	X51569			177	262				
39 skeletal muscle alpha-actin (b, f)	M20543	J00805	X05392		177	294	(c)	490		
40 extracell. signal-regulated kinase (f)	M84489		X59813				170	202		
41 alpha-tropomyosin	M18135	M69145			160	962				
42 myosin heavy chain	X62659	M26510			157	277				
43 cytoplasmic beta-actin (b,f)	X00351	X00182		M25013	156	592			104	592
44 HMG-2 (f)	M83665	M83235			155	378				
45 nucleolar protein NO38 / B23.1 (b)	J03969	X17200	X56039		153	193	(c)	307		
46 basic FGF receptor (b)	M37722	M24637	M55163		151	411	130	905		
47 octamer-binding protein (Oct-1) (b)	X13403	M29972	X57165		150	150	(c)	199		
48 nucleolin C23 (b)	M60858	X17199	X63091		149	392	0	425		
49 casein kinase II alpha subunit (f)	M55265	M59456			146	200				
50 GTP-binding protein (G-alpha-i1) (f)	X03642		X56089				145	1497		
51 non-muscle alpha-actinin (f)	X55187	M74143			139	537				
52 M-twist	M63650		M27730				134	698		
53 beta-nerve growth factor (b)	M35075	X04067	X55716		0	159	132	159		
54 transforming growth factor-beta 1	X02812	X12373			129	129				
55 Hox 7	M76732	X61922			122	368				
56 glycoprotein 96 (tra1) / 108K HSP	X15187	X04961			120	264				
57 Hox 2.6/XHox 1A	M36654		M26884				119	469		
58 tenascin	X56160	M23121			118	440				
59 prolyl 4-hydroxylase alpha subunit	M24486	M26217			114	1000				
60 fibronectin	X00739		M77820				110	674		
61 integrin (b)	X07979	M14049	M20140		108	1115	(c)	696		
62 neuropeptide Y (b)	K01911	M87295	M87296		105	173			(c)	173
63 CEF-10 / Cyr61	M32490	J04496			100	625				
64 B-creatine kinase (f)	X15334	M33714			99	423				
5'-HCR-containing genes										
65 insulin-like growth factor (IGF-I) (b)	M14155'	M32791	M29857	M32792	477	936	275	280	0	176
66 transforming growth factor-beta 1	X02812	X12373			437	447				
67 aortic smooth muscle alpha-actin	D00618	M13756			299	3357				
68 alpha-2 collagen gene type I	K01832	M25963			280	1487				
69 nuclear factor I-B	J04122	X51485			238	263				
70 fast myosin alkali light chain MLC1f	J05026	K02610			228	385				
71 Hox 5.1/CHox1.4	X17360	X52671			200	403				
72 c-myb	X16390	X12495			185	1051				
73 GABA-A receptor gamma-2 subunit	M62374	X54944			176	348				
74 Hox 2.3 (e)	X06762		X06592				169	812		
75 transforming growth factor-beta 3	M60556	X58127			159	2247				
76 c-jun	J04111	M57467			147	813				
77 cytoplasmic beta-actin (b)	M10277	X00182		M25013	147	325			0	305
78 alpha-tropomyosin	M15474	M69145			145	379				
79 Hox 3.3 PRII	X16511		X12500				128	128		
80 decorin/chondroitin	X53929	X63797			130	136				
81 B-creatine kinase	X15334	M33714			121	1148				
82 FGF receptor tyrosine kinase	X52832	M35196			119	160				
83 retinoic acid binding protein beta	Y00291	X57340			118	121				
84 ferritin H-subunit (b)	X03488	M16343	M15655		117	387	0	141		
85 alpha-1 type-III collagen	M26939	K01481			109	120				
86 Hox1.1 (e)	M17192	M59714			104	101				
genes with HCRs in introns										
87 c-fos first intron	K00650	M18043			118	438				

Note: A highly conserved region (HCR) is defined as a sequence with at least 70% similarity over 100 nt or more. (a) genes are sorted according to the HCR length. When the length of HCR is null (0), it means that we detected homology but that the conserved region did not fit the criteria of HCRs. The length of common non-coding regions sequenced in both species is indicated. (b) genes for which non-coding regions have been determined in at least three different vertebrate classes. (c) genes with no detectable similarity in non-coding regions. (d) the conserved region is found also in rat S-myc 3' UTR (Accession number: M29069). (e) 5'HCR from Hox1.1 and Hox2.3 are homologous. (f) there exists other from the same multigene family, very similar at the protein level (>80%) but that are totally divergent in their 3'NCRs. mamm. : mammals; amph. : amphibians.

unlikely that this conserved elements are protein encoding exons. We cannot exclude the presence of small exons such as the 14 nt long coding sequence in the 3'-UTR of SERCA2B that is used by alternative splicing for the expression of SERCA2A (fig. 1a); however these exons cannot account for the long HCRs found.

Conserved regions do not code for any known structural RNA

The non-coding regions of the HCR-containing genes were compared with all functional RNA genes described in GenBank (ribosomal-RNA, transfer-RNA, small cytoplasmic or nuclear RNA). We detected several non-coding sequences with significant homology to the small cytoplasmic RNA 7SL, but never within HCRs. The program tRNA-scan (17) also failed to predict any potential tRNA-encoding sequence. Thus, conserved regions do not correspond to any overlapping gene coding for a known functional RNA.

Distribution of highly conserved regions

In order to estimate the frequency of genes containing a HCR in their non-coding regions and study HCRs distribution within genes, we searched in GenBank for orthologous genes from species belonging to different classes of vertebrates. We chose all couples for which the complete coding part and at least 200 nt from the 3'-NCR, 5'-NCR or one complete intron was sequenced in both species. Table 2 gives the frequency of HCR-containing genes among orthologous genes with sequenced non-coding regions, for the different parts of genes and between mammals and birds, amphibians or fishes. These results are rough underestimates since it is possible that some of the genes in our list are not strictly orthologous or that they contain HCRs in a part of their NCRs that has not been sequenced; therefore they can only be used for comparison. We also calculated the average length of these HCRs and the ratio conserved region length over the total common length of sequence determined in both vertebrates classes.

The first observation is, as expected, that no significant homology can be detected in non-coding regions of most orthologous genes. However, HCRs are not exceptional: about 30% of genes contain in their 3'-NCR a HCR between mammals and birds, with an average length of 390 nt representing 17% of the total length of 3'-NCR sequenced in both classes.

Secondly, HCRs concern only limited parts of genes. The mutation rate is known to vary between genes from a same genome (18). This variation can be attributed to various factors such as structural features of the chromatin, replication or transcriptional activities that can affect the susceptibility to mutation or the efficiency of repair over DNA regions (19). Thus, we searched if the conservation that we observed between mammals and birds affects the whole gene. Among the 64 couples of genes that contain a 5'- or 3'-HCR and for which we could compare their coding part, we found only four cases with high similarity (>70%) at their fourfold degenerate sites. We had a sample of 34 orthologous genes with which we could test both the presence of 5' and 3' HCR; calculation of the table of contingency for the presence/absence of conserved region showed that 3'-HCRs are not associated with 5'-HCRs ($\chi_1{}^2=0{,}42$). Moreover, HCRs are generally only discrete portions of non-coding regions (see fig. 1). Therefore, the conservation that we observed does not concern the entire gene and cannot be attributed to some structural features of a chromosomal domain that would protect DNA from mutations.

Table 2. Frequency of HCR-containing genes, distributionand extent of HCRs within genes

	mammals versus:		
	birds	amph.	fishes
orthologous genes with >200 nt sequenced in 5'NCR	63	25	13
frequency of 5'HCR-containing genes (a)	17,4% (11)	4% (1)	0
total common sequenced length (nt)	49942	13998	6089
average 5'HCR length (nt)	221 ± 126	275	ND
5'HCR total length/total common sequenced length	4,90%	2%	0%
orthologous genes with >200 nt sequenced in 3'NCR	121	70	37
frequency of 3'HCR-containing genes (a)	29,7% (36)	17,1% (12)	5,4% (2)
total common sequenced length (nt)	80554	36888	18853
average 3'HCR length (nt)	390 ± 305	228 ± 106	254 ± 213
3'HCR total length/total common sequenced length	17,40%	7,40%	2,7%
orthologous genes with >200 nt sequenced in introns	19	6	5
frequency of intron-HCR-containing genes	5,2% (1)	0	0
total common sequenced length (nt)	25655	9699	4931
average intron HCR length (nt)	118	ND	ND
intron HCR total length/total common sequenced lengt	0,5%	0%	0%

Note: (a) because of the criteria that we have chosen (complete coding part, NCR > 200nt), some of the HCRs listed in table 1 are not included here. amph.: amphibians.

Table 3. Frequency of HCRs and classes of genes

mammals/birds orthologous genes	with 3'HCR	without 3'HCR
Total (a)	29,7% (36)	70,3% (85)
class of product		
DNA-binding protein or cytoskeletal protein	48,8% (21)	51,2% (22)
enzyme, hormone or hormone receptor	5,1% (2)	94,9% (37)
other (b)	33,3% (13)	66,7% (26)
Expression pattern (c)		
limited	19,4% (14)	80,6% (58)
wide	50% (19)	50% (19)

Note: (a) are only included genes with complete coding part and 3'NCR >200nt. (b) see text. (c) genes for which expression pattern could not be determined are not included.

Thirdly, we observed that HCRs are not evenly distributed in the gene: they are predominantly found in the 3' NCRs, where they are much longer than those in 5'-NCRs (table 2). In a significant number of cases (>10%), the similarity in 3'-HCRs is even higher than in the coding part of the gene. There is a bias in the database since the number of 3'-NCRs longer than 200 nt sequenced in two different classes is almost twice as for 5'-NCRs. However, the average common sequenced length in 5'- and 3'-NCRs is approximately the same (respectively 792 and 665 nt). Therefore, this bias cannot account for the difference in frequency of 5' and 3' HCR-containing genes (17% versus 30%) nor for the difference between the ratio conserved/total common sequence length (5% versus 17%). HCRs appear to be even scarcer in introns (1 HCR found among 19 genes). This means that there is generally a stronger selective pressure to preserve the sequence of 3'-NCRs than of 5'-NCRs or introns. This finding is unexpected because it is generally thought that the most important regulatory non-coding region is the promoter.

One other notable feature of 3'-HCRs is that they are always located in the transcribed part of 3'-NCRs. We did not find any significant conserved region in 3'-flanks, and among the 13 couples of 3'-HCR-containing genes for which at least 100 nt of the 3'-flank had been sequenced in two species (390 nt in average), the conserved regions never extended more than 30 nt after the polyadenylation site. Together with the fact that HCRs are very rare in introns, this indicates that there exists a constraint that operates on the mature mRNA and not at the pre-mRNA or DNA levels. Thus we can infer that the 3'-HCRs are involved in post-transcriptional mechanisms.

Evolution of highly conserved regions

HCRs are detected even between the most distantly related classes of vertebrates (except agnatha that are poorly represented in GenBank): the Na+/K+-ATPase beta-subunit gene contains in its 3'-UTR a HCR of 400 nt between mammals and osteochthyes that diverged more than 420 My ago (fig. 1a, [8]). However, table 2 shows that conserved regions are scarcer and, in average, smaller as species are more distantly related. To confirm this result, we searched in our list for HCR-containing genes that have been sequenced in at least 3 different vertebrate classes. In 12 cases among 18, homology was detected between all classes with the strongest conservation observed between mammals and birds, except for beta nerve growth factor [53] and alpha 1 type II collagen [25]; in the 6 other cases, homology was detected between mammals and birds but not between more distantly related classes. This can be explained by the fact that similarity is already weak between mammals and birds and cannot be detected anymore with fishes or amphibians (e.g. neuropeptide Y [62]). An alternative explanation can be that genes are not really orthologous: for example a duplication of the skeletal muscle alpha-actin ancestor gene arose in the amphibian lineage, leading to at least two distinct skeletal isotypes that are not strictly comparable to their single mammalian counterpart (20) [39]. All these results show that HCRs are slowly evolving sequences that diverge from a common ancestor and allow to reject the hypothesis that they are due to virus-mediated horizontal gene transfer or to experimental artefact such as recombination with contaminant DNA during cloning.

Many of the HCR-containing genes are known to belong to multigene families that arose by successive duplications of ancestor genes. Interestingly, whereas related genes sometimes code for isoforms that are very similar at the protein level, we found that their non-coding sequences are specific of each isotype (except for N- and S-myc [5], Hox1.1 and Hox2.3 [74, 86]) (see table 1). The most striking examples are the calmodulins and replacement variant histones H3.3. Mammalian H3.3A and H3.3B genes code for the same protein and yet have distinct 5' and 3' regions. In both genes the entire 3'-UTRs (>500 nt) remained highly conserved (>80% similarity) since mammals and birds divergence (21, 3) [9, 12]. This multigene one-protein principle is also observed for vertebrate calmodulins: in mammals and toleost fishes there are respectively three and four distinct genes that code for exactly the same protein (22, 23). Of these, there is at least one for which the 3'-UTR has been exceptionally conserved (80% similarity over 550 nt) between mammals, birds and amphibians [10]. If these genes were truly redundant, one should expect that some of the copies would have been lost during such a long period of evolution. If the presence of multiple copies arose from necessity of high protein production, one should predict the conservation of the number of copies but not of each one specifically, particularly because mechanisms of conversion between homologous sequences are relatively frequent in vertebrates (24). Therefore, these multiple genes are probably maintained because each one has a specific expression in response to different stimuli. The presence of specific 3'-HCR in such genes is thus particularly meaningful and suggests a role of 3'-HCRs in the unique pattern of expression of each member of multigene families.

3′-HCRs are preferentially found in genes that are essential to cell life

In order to get insight on the potential function of conserved sequences, we searched if the presence of HCRs was associated with a particular expression pattern. We classified HCR-containing genes in two groups: genes expressed in a wide range of tissue and genes with limited or tissue-specific expression. Information about the expression of genes was extracted from OMIM database (Victor A.McKusik's catalogue of genes, accessed through GDB, 25) or from papers given in reference in GenBank. The most important feature is that 3′-HCRs are clearly associated with widely expressed genes ($\chi_1^2=11.06$) (table 3). Interestingly this bias is not observed with 5′-HCRs that are found at about equal frequency in genes with limited or wide expression ($\chi_1^2=0.39$). Thanks to the numerous 3′-HCR-containing genes that we found it was possible to study their distribution according to the function of their product. Genes were classified in the seven most representative groups: enzymes, DNA-binding proteins, cytoskeletal proteins, extra-cellular matrix proteins, hormones, hormone receptors, storage or transport proteins. There again we found a striking difference: 3′-HCRs occur predominantly in genes encoding DNA-binding proteins or cytoskeletal proteins, about 10 times more frequently than in genes encoding enzymes, hormones or hormone receptors (table 3). In total, 75% (12/16) of genes that are widely expressed and that encode DNA-binding proteins or cytoskeletal components contain a 3′-HCRs.

The basal cellular mechanisms are probably similar in all vertebrates whereas differences appear at higher level of organisation (tissues, organs…). It is thus not surprising that the regulation of genes involved, for example, in the hormonal system may have radically changed since the divergence of vertebrates, whereas mechanisms of regulation of the genes that are essential to cell life, have remained the same. What these results show is that these latter mechanisms of regulation often involve the 3′-UTR and operate at the post-transcriptional level.

Compositional properties of HCRs

In order to see if HCRs have particular structural properties, we compared the nucleotide composition of conserved and non-conserved regions. We observed that while 5′-HCRs have the same composition as non-conserved regions that surround them, 3′-HCRs are strongly enriched in A+T; not only are they richer than 3′-NCRs that do not contain HCR (64% versus 54% A+T in mammals and birds), but also than non-conserved regions that surround them (58% A+T). This difference is mostly due to an excess of T (U) in the transcribed strand (34% in 3′-HCRs versus 27% in 3′-NCRs that do not contain HCR). Mouchiroud *et al.* (15) have shown that the G+C-content of each part of genes is subject to high scale compositional constraints that affect large chromosomal domains (> 100 kb). These long regions of either high, medium or low G+C-content are termed isochores (26) (respectively H3, H1+H2, L1+L2). We observed that 3′-HCRs are found in G+C-rich as well as in G+C-poor isochores. However, whereas the A+T content of non-conserved 3′-NCRs varies according to the isochore (from 46% in average in H3 to 64% in L1+L2), 3′-HCRs remain A+T rich in all isochores (64–65%). These results suggest that the A+T-richness of 3′-HCRs corresponds to a functional requirement.

Potential functions for HCRs located in introns or 5′ non-coding regions.

We have found only one HCR in introns, among 19 couples of orthologous genes with their introns (complete and > 200 nt) known in two vertebrate classes (table 2). It is located in the first intron of c-fos [87] and its role is unknown. Whereas the presence of introns has been shown to be necessary for the correct processing or export of certain mRNAs (27), these limited data suggest that their primary sequences are generally not subject to strong constraints.

The observation of the profiles of similarity in the 5′ NCRs reveals two areas that are preferentially conserved: sequences surrounding the transcription start-site and sequences of the 5′-UTR just upstream the initiation codon. This pattern strongly suggest a role in the regulation of transcription and/or translation.

The creatine kinase B, c-jun (fig. 1b) and actins genes [81, 76, 67, 77] are typical examples where a HCR is found to encompass the CAAT- and TATA-boxes and hence are very likely to play a role in the regulation of transcription. As we discussed above, HCRs are much longer than sequences that are necessary to specify the binding of a single protein. However, formation of the transcription initiation complex requires the concerted binding of numerous factors on adjacent DNA sites, which could explain the maintenance of a short HCR.

Among genes that contain a HCR in their 5′-UTR, we found two examples where the conserved element has been shown to play a role in regulation of translation: transforming growth factor beta-3 (28) [75] and ferritin heavy chain (fig. 1b, [84]). Translational regulation of the ferritin mRNA has been thoroughly studied (reviewed in 29). The 5′-UTR of this mRNA contains a sequence called Iron Response Element (IRE) that adopts a stem-loop structure which is recognised by a specific cytoplasmic binding protein known as IRE-BP. In response to iron starvation, the IRE-BP is activated and binds with high affinity to this IRE thus repressing the translation of this protein. The IRE is about 30 nt long, whereas the conserved region covers 120 nt. The role of sequences flanking the IRE remains to be determined.

These observations are indeed not very surprising: 5′-HCRs can correspond to the presence of several adjacent signals recognised by specific nuclear DNA-binding proteins, or to an RNA region that adopts a particular spatial structure interacting with a cytoplasmic RNA-binding protein. The patterns of conservation of most of the 5′-HCRs that we found can be explained by one and/or the other of these two models. However, 5′-HCRs are longer than currently characterised regulatory elements, thus it would be interesting to determine experimentally the precise function of each part of conserved regions.

Role of 3′-HCRs in post-transcriptional mechanisms of regulation

We have shown that 3′-HCRs are probably involved in post-transcriptional mechanisms that require A+T-rich regions in the 3′-UTR of mRNAs. To examine the potential role of 3′-HCR in the different post-transcriptional steps (mRNA nucleo-cytoplasmic export, cytoplasmic localisation, translation and degradation), we reviewed the papers given in reference in the sequence entries. We found several well studied examples where we could associate sequence conservation with a particular function. However, none of them is sufficient to explain all the 3′-HCRs that we characterised. The potential role of 3′-HCRs in each post-transcriptional step is discussed below.

mRNA degradation. Regulation of mRNA degradation is an important process in controlling the expression of a number of proteins. Interestingly, numerous studies have clearly demonstrated the role of 3′-UTR in the regulation of mRNA turnover (reviewed in 30,31). We found two genes with a 3′-HCR encompassing cis-acting elements that have been shown to influence mRNA stability: c-fos [7] and transferrin receptor (TfR) [16]. The c-fos 3′ mRNA destabilizing element has been thoroughly studied, and was recognised as centred around a 75 nt A+T-rich sequence (32). However, the c-fos 3′-HCR covers more than 650 nt and thus cannot be fully explained by this relatively short signal (see fig. 1a). The best understood case is the TfR mRNA (reviewed in 29). The cis-acting regulatory sites within the 3′-UTR have been mapped by detailed deletion and mutation analysis; they correspond to two A+T-rich HCRs of respectively 200 and 250 nt separated by a dispensable segment whose length has not been conserved during evolution (see fig. 1a). The first HCR contains 2 IREs (the same structure as the one found in the 5′-UTR of ferritin mRNA described above) and a stem loop structure, while the second encompasses 3 additional IREs (33). In cells grown in a medium with high iron content, the stem-loop in the first HCR confers a high turn-over rate to the TfR mRNA, whereas in an iron-depleted medium IRE-BPs are activated and bind at least four of the five IREs in a 1:1 stoichiometry, thus protecting TfR mRNA against degradation (34, 35).

This example is interesting in clearly demonstrating the link between a long highly conserved non-coding region and its function in post-transcriptional regulation through the formation of adjacent small stem loop structures. Several data suggest that many of the 3′-HCRs that we found may also correspond to cis-acting determinant of mRNA stability. First, several other 3′-HCR-containing genes are known to be regulated at the level of mRNA stability: N-myc (36) [5], ornithine decarboxylase (ODC) (37) [23], cyr-61 (38) [63], c-myb (39) [38]. Furthermore, as already mentioned, many of the genes that we found code for regulatory proteins which mRNAs are expected to have a short half-life in order to allow a fine tuning of their expression. Finally, the common feature of the cis-acting destabilizing elements that have been characterised to date is their A+T-richness (30), which is also a feature shared by most of 3′-HCRs.

It is however clear that this model cannot account for all the 3′-HCRs that we detected. First, as pointed out by Lemaire *et al.* (2), it is not expected that the very large dystrophin gene [2], which transcription takes about 20 hours, is regulated at the level of mRNA stability. Furthermore, many of the 3′-HCR-containing genes encode cytoskeletal proteins which mRNAs are known to be very stable (half life > 4h: fibronectin [60], integrin [61], tropomyosin [41] (40); >24h: beta-actin [43], neurofilament [15] (41)). Therefore, it is likely that other post-transcriptional processes are responsible for these 3′-HCRs.

mRNA transport and localisation. Recent experiments have demonstrated that many mRNA have a non-uniform distribution in the cytoplasm. This specific subcellular localisation of mRNA provides a mechanism for protein targeting within developing or differentiating cells and thus participates to the establishment of cellular morphology (reviewed in 42). mRNA sorting is an active process that involves the cytoskeleton (43). In all cases studied to date, this process is mediated by cis-acting sequences in the 3′-UTR (44). Interestingly, it has been shown that 53 nt from the proximal portion of the 3′-UTR of the beta-actin gene were sufficient to confer specific subcellular localisation to an otherwise non-localised reporter transcript (44). Unfortunately, the authors have not published the exact position of this signal; it would be meaningful that it resides in the HCR found at the beginning of the 3′-UTR [43].

The mRNAs of two other 3′-HCR-containing genes have been shown to be localised in differentiated cells: vimentin (45) [26] and myosin heavy chain (46) [42]. These examples support the hypothesis that 3′-HCRs may also be involved in the control of mRNA localisation. It would be interesting to test this model with the numerous 3′-HCR-containing genes that encode cytoskeletal proteins which localisation is probably essential for control of cellular morphology. Moreover, this hypothesis suggests a possible role for the existence of multiple genes coding for the same protein: their specific 3′-UTR may target each mRNA to a particular cellular localisation, thus giving a specific destination to the protein. Thus, the level at which each isotype is expressed would participate to the establishment of cellular morphology.

mRNA translation. Another potential role for 3′-HCR is the control of mRNA-translation. It has been reported that A+T-rich sequences in 3′-UTRs such as the one that confer instability to c-fos mRNA, have also an inhibitory effect on translation in *Xenopus* oocytes (47). The mechanisms by which 3′-UTR influence translation are not understood, and thus the link between sequence conservation and function remains obscure. The ODC gene has an A+T-rich, highly conserved 3′-UTR [23] and, as already mentioned, its expression is regulated at the level of mRNA stability. Interestingly, it has also been shown that its 3′-UTR strongly influences the efficiency of translation (48), which raises the possibility that a HCR might correspond to several adjacent signals involved in different mechanisms.

mRNA nucleo-cytoplasmic export. Finally, though poorly documented, we must mention the potential role of 3′-HCRs in the nucleo-cytoplasmic export of mRNA. This is an active process that involves the poly(A) tail (49). Interestingly, recent experiment have shown that a protein, named A+U binding factor (AUBF), accelerates the nuclear export of mRNA containing an A+T rich sequence in their 3′-UTR (50).

CONCLUSIONS

We have shown that non-coding sequences can be subject to a strong selective pressure. Interestingly, whereas most attention is generally focused on the control of transcription, we found that selective constraints are stronger on transcribed 3′-NCRs than on any other non-coding regions, even promoter sequences. This stresses the importance of post-transcriptional events in control of gene expression. Our results suggest that such mechanisms have been conserved during vertebrates evolution principally for genes that are essential to cell life. The control of these post-transcriptional mechanisms is so important that in some cases, multiple genes encoding a same protein have been maintained in order to allow a response to various stimuli or to target mRNA to specific localisation through different 3′-UTRs.

However, whereas our work allowed to reveal non-coding regions of functional importance, very little is known about the precise mechanisms in which they are involved. None of the different post-transcriptional steps can account for all the 3′-HCRs that we found. The A+T-richness of 3′-HCRs probably reflects a functional requirement, but this does not give much information since A+T-rich sequences in 3′-UTR have been shown to be involved in many different processes (see above). The only well understood case is the TfR mRNA 3′-UTR that contains six stem-loop structures playing a role in the control of its stability. Even though, the total length of these six elements is about 200 nt whereas the conserved region spans 450 nt. The role of theses additional conserved sequences is unknown. The length of the other known cis-acting elements (c-fos destabilizing sequences, beta-actin localiser signal) is under 75 nt. This contrasts with the average length of mammalians/birds 3′-HCRs which is near 400 nt. What mechanism(s) can necessitate conserved regions of more than 700 nt found in the 3′UTRs of N-myc, NCAM, BTG1, dystrophin and SERCA2B genes? Do long HCRs correspond to adjacent and independent short signals or to complex structures involved in a single process? Whereas we are familiar with the mechanisms of transcription or translation that account for the patterns of conservation that we observed in 5′-NCRs, there is clearly a gap between our current knowledge of post-transcriptional mechanisms of regulation and the constraints that we characterised on 3′-UTR.

REFERENCES

1. Yaffe,D., Nudel,U., Mayer,Y., and Neuman,S. (1985) *Nucleic Acid Res.* **13**, 3723–3737.
2. Lemaire,C., Heilig,R. and Mandel,J.L. (1988) *EMBO J.* **7**, 4157–4162.
3. Hraba-Renevey,S. and Kress,M. (1989) *Nucleic Acid Res.* **17**, 2449–2461.
4. Kajimoto,Y. and Rotwein,P. (1991)*J. Biol. Chem.* **266**, 9724–9731.
5. Rouault,J.P., Samarut,C., Duret,L., Tessa,C., Samarut,J. and Magaud,J.P. (1993) *Gene* In Press.
6. Marzluff,W.F. and Pandey,N.B. (1988)*Trends Biochem. Sci.* **13**, 49–52.
7. Li,W.H., Luo,C. and Wu,C. (1985) In MacIntyre,R.J. (ed.), Molecular Evolutionary Genetics. Plenum Press, New York, pp.1–94.
8. Benton,M.J. (1990) *J. Mol. Evol.* **30**, 409–424.
9. Goodman,M., Czelusniak,J., Koop,B.F., Tagle,D.A. and Slightom,J.L. (1987) *Cold Spring Harbor Symposia on Quantitative Biology* **LII**, 875–890.
10. Burks,C., Cassidy,M., Cinkowsky,M.J., Cumella,K.E., Gilna,P., Hayden,J.E.D., Keen,G.M., Kelley,T.A., Kelly,M., Kristofferson,D. and Ryals,J. (1991) *Nucleic Acid Res.* **19**, 2221–2225.
11. Gouy,M., Gautier,C., Attimonelli,M., Lanave,C. and Di Paola,G. (1985) *CABIOS* **1**, 167–172.
12. Altschul,S.F., Gish,W., Miller,W., Myers,E.W., and Lipman,D.J. (1990) *J. Mol. Biol.* **215**, 403–410.
13. Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) *CABIOS* **8**, 189–191.
14. Jacobzone,M. and Gautier,C. (1986) *Publication interne, Laboratoire de Biometrie, Universite Claude Bernard—Lyon I.*
15. Mouchiroud,D., D'Onofrio,G., Aissani,B., Macaya,G., Gautier,C. and Bernardi,G. (1991) *Gene* **100**, 181–187.
16. Beckmann,J.S. and Weber,J.L. (1992) *Genomics* **12**, 627–631.
17. Fichant,G.A. and Burks,C. (1991) *J. Mol. Biol.* **220**, 659–671.
18. Wolfe,K.H., Sharp,P.M. and Li,W.H. (1989) *Nature* **337**, 283–285.
19. Boulikas,T. (1992)*J. Mol. Evol.* **35**, 156–180.
20. Mohun,T., Garrett,N., Stutz,F. and Spohr,G. (1988) *J. Mol. Biol.* 202, 67–76.
21. Wells,D., Hoffman,D. and Kedes,L. (1987) *Nucleic Acid Res.* **15**, 2871–2889.
22. Nojima,H. (1989) *J. Mol. Biol.* **208**, 269–282.
23. Matsuo,K., Sato,K., Ikeshima,H., Shimoda,K. and Takano,T. (1992) *Gene* **119**, 279–281.
24. Murti,J.R., Bumbulis,M., and Schimenti,J.C. (1992) *Mol. Cell. Biol.* **12**, 2545–2552.
25. Pearson,P.L., Matheson,N.W., Flescher,D.C. and Robbins,R.J. (1992) *Nucleic Acid Res.* **20** (supplement), 2201–2206.
26. Bernardi,G. (1989) *Annu. Rev. Genet.* **23**, 637–661.
27. Huang,M.T.F. and Gorman,C.M. (1990) *Nucleic Acid Res.* **18**, 937–947.
28. Arrick,B.A., Lee,A.L, Grendell,R.L. and Derynck,R. (1991) *Mol. Cell. Biol.* **11**, 4306–4313.
29. Kühn,L.C. and Hentze,M.W. (1992) *J. Inorg. Biochem.* **47**, 183–195.
30. Peltz,W., Brewer,G., Bernstein,P., Hart,P.A., and Ross,J. (1991) *Crit. Rev. Euk. Gene Express.* **1**, 99–126.
31. Hentze,M.W. (1991) *Biochem. Biophys. Acta* **1090**, 281–292.
32. Shyu,A., Greenberg,M.E. and Belasco,J.G. (1989) *Gene Develop.* **3**, 60–72.
33. Chan,L.L., Grammatikakis,N., Banks,M. and Gerhardt,E.M. (1989) *Nucleic Acid Res.* **17**, 3763–3771.
34. Koeller,D.M., Casey,J.L., Hentze,M.W., Gerhardt,E.M., Chan,L.N., Klausner,R.D. and Harford,J.B. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 3574–3578.
35. Müllner,E.W. and Kühn,L.C. (1988) *Cell* **53**, 815–825.
36. Babiss,L.E. and Friedman,J.M. (1990) *Mol. Cell. Biol.* **10**, 6700–6708.
37. Watson,G. and Paigen,K. (1988) *Mol. Cell. Biol.* **8**, 2117–2124.
38. O'Brien,T.P., Yang,G.P., Sanders,L. and Lau,L.F. (1990)*Mol. Cell. Biol.* **10**, 3569–3577.
39. Thompson,C.B., Challoner,P.B., Neiman,P.E. and Groudine,M. (1986) *Nature* **319**, 374–380.
40. Ryseck,R.P., Macdonald-Bravo,H., Zerial,M. and Bravo,R. (1989) *Exp. Cell Res.* **180**, 537–545.
41. Schwartz,M.L., Shneidman,P.S., Bruce,J. and Schlaepfer,W.W. (1992) *J. Biol. Chem.* **267**, 24596–24600.
42. Stewart,O. and Banker,G.A. (1992) *Trends Neurosci.* **15**, 180–186.
43. Singer,R.H. (1992) *Curr. Opin. Cell. Biol.* **4**, 15–19.
44. Kislauski,E.H. and Singer,R.H. (1992)*Curr. Opin. Cell. Biol.* **4**, 975–978.
45. Lawrence,J.B. and Singer,R.H. (1986) *Cell* **45**, 407–415.
46. Pomeroy,M.E., Lawrence,J.B., Singer,R.H. and Billings-Gagliardi,S. (1991)*Dev. Biol.* **143**, 58–67.
47. Kruys,V., Marinx,O., Shaw,G., Deschamps,J. and Huez,G. (1989) *Science* **241**, 852–855.
48. Grens,A. and Scheffler,I.E (1990) *J. Biol. Chem.* **265**, 11810–11816.
49. Schröder,H.C., Bachmann,M., Diehl-Seifert,B. and Müller,W.E.G. (1987) *Prog. Nucl. Acid Res. Mol. Biol.* **34**, 89–142.
50. Müller,W.E.G., Slor,H., Pfeifer,K., Hühn,P., Bek,A., Orsulic,A., Ushijima,H. and Schröder,H.C. (1992) *J. Mol. Biol.* **226**, 721–733.