Available online at www.sciencedirect.com

SCIENCE DIRECT

Biochimica et Biophysica Acta 1680 (2004) 24–33

BIOCHIMICA ET BIOPHYSICA ACTA BBA

http://www.elsevier.com/locate/bba

Genomic organization of the cadmium-inducible tandem repeat 25-kDa metallothionein of the oligochaete worm *Enchytraeus buchholzi*☆

Hans-Peter Schmitt-Wrede[a], Heike Koewius[a], Steffen Tschuschke[a], Hartmut Greven[b], Frank Wunderlich[a,*]

[a]*Division of Molecular Parasitology and the Centre of Biological and Medical Research, Heinrich-Heine University, Universitätsstr. 1, 40225 Düsseldorf, Germany*

[b]*Institute of Zoomorphology and Cell Biology, Heinrich-Heine University, Universitätsstr. 1, 40225 Düsseldorf, Germany*

Received 10 December 2003; received in revised form 27 July 2004; accepted 26 August 2004
Available online 11 September 2004

Abstract

The terrestric oligochaete worm *Enchytraeus buchholzi* survives in cadmium (Cd)-polluted environments by aid of its Cd-inducible 25 kDa cysteine-rich protein (CRP). Here, we analyze promoter and structure of the *crp* gene and compare its relationship to MT genes. The *crp* gene, approximately 12 kbp long, consists of 10 exons with exons 2 to 9 encoding eight almost identical repeats of predominantly 31 amino acids of the CRP. The introns of the *crp* gene contain various repetitive elements including retrotransposon-like sequences. The 683-bp promoter of the non-constitutive *crp* gene exhibits a much higher basal activity than the mouse MT-II promoter in HepG2 cells. Essential for *crp* promoter activity is the distal region (−683/−521) with a GC box and the proximal region (−308/−8) with the four MREa, b, c, d and AP-1, -2, -3 elements, whereas the central portion (−521/−309) with CAAT box, CRE and a XRE causes promoter repression. The TATA box-, MREc- and the AP-2, -3-containing region are required for high *crp* promoter activity. Our data support the view that the *crp* gene is a unique MT-gene and has evolved by exon duplications from a MT-like ancestral gene.

Keywords: Cadmium-resistance; Cysteine-rich protein; *Enchytraeus buchholzi*; Metal-responsive element; 25-kDa metallothionein

1. Introduction

The anthropogenic pollution of the environment by heavy metals is not only a serious hazard for animals, plants and ecosystems, but also for human health. Anthropogenic sources are considered to cause, for example, more than 90% of the cadmium (Cd) input into biosphere of approximately 30,000 tons/year [1]. This metal is increasingly mobilized due to sustained acidification of the soil caused by acid rain, which in turn increases Cd-bioavailability and, thus, Cd-toxicity [2]. Cd is highly toxic due to its strong affinity to several ligands such as purines, pyrimidines, phosphates, porphyrines, cysteine- and histidine-residues of proteins [3,4].

The toxicity of Cd can be partly reduced in many organisms through metallothionein (MT) (recent reviews Refs. [5–8]). Transition metals and various other stressors induce the biosynthesis of these 6–7-kDa proteins [9]. MTs are encoded by genes of diverse exon/intron structure in invertebrates and a more homogeneous 3 exon/2 intron structure in vertebrates [10,11]. In general, MTs are predominantly regulated at the transcriptional level [12]. Metal-responsive elements (MREs) are common to almost all

Abbreviations: Cd, cadmium; MT, metallothionein; CRP, cysteine-rich protein; HSE, heat shock element; Inr, initiator element; MRE, metal responsive element; LTR, long terminal repeat; LINE, long interspersed nucleotide element; kbp, kilo base pairs; SEAP, secreted form of human placental alkaline phosphatase

☆ The nucleotide sequences in this paper have been submitted to the DDBJ, EMBL, GenBank databases with accession numbers AJ565921, AJ565922.

* Corresponding author. Tel.: +49 211 811 3401; fax: +49 211 811 4734.
E-mail address: frank.wunderlich@uni-duesseldorf.de (F. Wunderlich).

0167-4781/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbaexp.2004.08.007

known MT promoters and represent the essential *cis*-acting elements for metal induction [13–15].

Besides MTs, there have been also described, though very scarcely, larger Cd-binding proteins of higher molecular mass, which are also capable of detoxifying Cd. In particular, such proteins have been reported to occur in several invertebrates, but their molecular characterization has been largely neglected to date [16]. For instance, the cysteine-rich 25-kDa CRP protein of the terrestric oligochaete worm *Enchytraeus buchholzi* contributes to the capability of the worms to survive in Cd-polluted environments [17–19]. Even in yeast, the CRP protein is efficient to mediate Cd-resistance. This becomes evident as an acquisition of high Cd-resistance of Cd-hypersensitive yeast cells after being transformed with *crp* [20]. The amino acid sequence of the CRP protein has been deduced from cDNA analysis and reveals a hitherto unique arrangement of eight tandem repeats. The repeats are 31-amino-acid long and contain nine cysteines, six of them are arrangeds in Cys-X-Cys and Cys-Cys segments. These are obviously important for coordination of Cd, similar as in MTs [20–23].

There is circumstantial evidence that the *crp* gene is more tightly regulated at the transcriptional level than MT genes. Indeed, the *crp* gene is not constitutively expressed as MT genes. Rather, its expression is induced specifically by Cd whereas Zn is clearly a weaker inducer than Cd at the highest sublethal metal-concentration. Other stressors such as Cu, Pb, H_2O_2 and heat do not induce *crp* transcription [19]. This study investigates the structure and promoter of the *crp* gene. Comparison with MT suggests that the *crp* gene can be considered as a member of the large MT gene family and is derived by exon duplication from an ancestral MT gene.

2. Materials and methods

2.1. Cultivation of enchytraeids

Mass cultures of *E. buchholzi* were grown in artificial soil as detailed previously [18]. For Cd-exposure, mature worms were kept in a fluid medium and then exposed to 3 mg Cd/l at 20 °C for 3 days as described elsewhere [19].

2.2. Genomic library screening

To isolate genomic sequences of *crp*, a partial genomic library of *E. buchholzi* in λZAPII (Stratagene, Heidelberg, Germany) was screened with [^{32}P]-labelled probes synthesized by random priming from the full length crp cDNA. Plaque hybridization was carried out as detailed elsewhere [18]. A 1861-bp genome clone (CRP-Lambda) was isolated.

2.3. Cloning of crp genomic fragments

The 5′ flanking region, the 3′ region, exon 6 and parts of intron 6 of the *crp* gene were cloned using the polymerase chain reaction (PCR)-based method for walking in uncloned genomic DNA as described recently [24]. Genomic DNA was digested with *Bsh*1236I, *Dra*I, *Ecl*136II, *Eco*47III, *Eco*RV, *Pvu*II, *Rsa*I, *Sca*I, *Sma*I, *Ssp*I and *Stu*I, before a special adaptor was ligated to the DNA fragments (adaptor oligonucleotides: 5′-CTAATACGACTCACTCACTATAGGGC-TCGAGCGGCCGCCCGGGCAGGT-3′; PO_4-5′-ACCTGCCCA-3′-NH_2). For the primary amplification with the Expand High-Fidelity PCR System (Roche Molecular Biochemicals, Mannheim), *crp*-specific primers (CRP1: 5′-TGCTTCAGCATGAGCGTGTGATGCCATGAT-3′; GPCR-4: 5′-TGCACCTCGTTACATCTCTGATAT-CAGTTA-3′; GPCR-6: 5′-CTCAATATCATGCTAGTG-TAGTGGATAGCACTGG-3′) and an adaptor primer AP1 (5′-GGATCCTAATACGACTCACTATAGGGC-3′) were used for 35 cycles (15 s, 94 °C; 6 min, 68 °C) with a final extension step (15 min, 72 °C). A second nested PCR was performed with *crp*-specific primers (CRP2: 5′-GTG-TGATGCCATGATTATTTTTCCACTTGAAGTCT-3′; GPCRcDNA+1036: 5′-GTTTCTGAAGTGCTCCTAC-TATTCACTGAC-3′; GPCR-7: 5′-TGTTTCAGTTGAA-GATTGCCATGTGGTCC-3′) and the adaptor primer AP2 (5′-AATAGGGCTCGAGCGGC-3′) for 30 cycles under the same conditions. The *crp*-specific primers were derived from the *crp* cDNA (GenBank accession number X79344).

For genomic PCR amplification, the following oligonucleotides based on the cDNA sequence and intronic gene-fragments were used: 5′-AATCAACCAATCGCA-GAACATCTAACAAAA-3′ (GPCR-1); 5′-ACAAGAA-TAAACTTACTGTCGCACTTAGGA-3′ (GPCR-2); 5′-TCCTAAGTGCGACAGTAAGTTTATTCTTGT-3′ (GPCR-3); 5′-GTGCAAGTTACTGCTCAGTGATT-CTGAATT-3′ (GPCR-5); 5′-GGGCGATCAGTT-CCCAGTCCAACATTAGTCG-3′ (GPCR-8); 5′-CTCACACTTGCAGCTTGGACCACATGGGCA-3′ (CRPcDNA-573); 5′-TTGTTGGCTGCCCATGCGGGT-CACAGTGCAC-3′ (CRPcDNA+629); 5′-AAACAATC-CATCACAGAATCAATAAAAGATAA-3′ (CRPcDNA-1386). Polymerase chain reactions were performed for 30 cycles with the Expand High-Fidelity PCR System using 200 ng of DNA. PCR products were cloned in the *pcDNA 2.1* vector using the TOPO TA Cloning kit (Invitrogen, Groningen, NL) or in *pMOSBlue T*-vector (Amersham-Pharmacia, Freiburg, Germany) for sequence analysis [25].

2.4. Promoter constructs

A 944-bp genomic DNA fragment containing the 5′ untranslated region and the 5′ flanking region of the *crp* gene served as a template for promoter deletion constructs. DNA fragments with various lengths of the 5′ flanking region were generated by PCR using the following gene specific primers: −689*Xho*I (5′-TCTCGAGTAACCGATACCGCCTCAATAATAT-3′), −683*Xho*I (5′-ACTCGAGGATACCGCCTCAATAA-

TATG-3′), −566*Xho*I (5′-CCGCTCGAGTGATTTTAGAGTAAAGAATT-3′), −521*Xho*I (5′-CCGCTCGAGATTTACAATTTTGAAGTATG-3′), −489*Xho*I (5′-CCGCTCGAGTTTTAGTCTACGGAGTTCT-3′), −421*Xho*I (5′-CCGCTCGAGTTAAACCGATTAAAAGTTTG-3′), −376*Xho*I (5′-CCGCTCGAGTTGTCATGGACTACATTATA-3′), −308*Xho*I (5′-CCGCTCGAGATAAGTCTTCTGTAAACTGC-3′), −165*Xho*I (5′-CCGCTCGAGGTTAAACTCTCAGCTATTCT-3′), −132*Xho*I (5′-CCGCTCGAGCTAAATGTCGTGTGCAGA-3′), −101*Xho*I (5′-CCGCTCGAGATGTTACAATTATTCTGACT-3′), −211comp*Hin*dIII (5′-AAGCTTAGTCCCCTTTATACCTTATCAAAACGATCTTCA-3′), −26comp*Hin*dIII (5′-CCCAAGCTTATATTCTCTGGATGTAAACA-3′), −8comp*Hin*dIII (5′-CCCAAGCTTCAAGTTCCATTTGTTTATAT-3′). A 410-bp DNA fragment of the mouse metallothionein II promoter (mMTII) (GenBank accession number K02236) was generated by genomic PCR. Genomic amplification (30 cycles, 1 min 95 °C, 1 min 55 °C, 2 min 72 °C) was performed with *Pfu*-DNA-Polymerase (Promega, Mannheim, Germany) and 450-ng genomic mouse DNA using the primers MT-IIX (5′-CCGCTCGAGGCGGGAAGCAAGCAGGTGT-3′) and MT-IIH (5′-CCGAAGCTTCGGAGGCACAGGAGCAGTT-3′). All promoter fragments were ligated to the *Xho*I/*Hin*dIII cloning sites of the vector *pSEAP-Enhancer* (Clontech, Heidelberg, Germany). All constructs were analyzed by sequencing of both DNA strands.

2.5. RNase protection assay

Poly $(A)^+$-RNA was isolated from Cd-treated earthworms. To produce an anti-sense probe, a 944-bp promoter fragment in vector *pMOSBlue* was linearized with Eco RV and transcribed with T7 RNA Polymerase (MAXIscript, Ambion, TX, USA) and 50 μCi [α-^{32}P]-dUTP (>22.2 TBq/mmol, ICN, Eschwege, Germany). The protection assay was performed according to guidelines of the RPAII Kit (Ambion). In brief, the gel purified probe (8×10^4 counts per minute) was hybridized overnight with 2-μg poly $(A)^+$-RNA at 45 °C and unhybridized RNA was digested subsequently with a RNase A/T_1 mix. The protected fragments were separated on a 6% sequencing gel. The 944-bp promoter fragment was radioactively sequenced and used as a size marker.

2.6. Transfection of cells and reporter gene assays

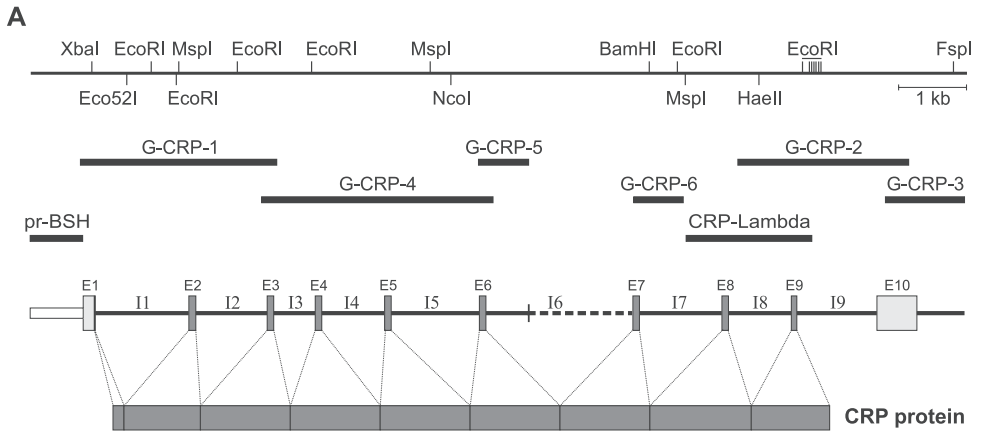
The human hepatoblastoma cell line HepG2 (ATCC HB-8065) was maintained in RPMI1640 (Invitrogen, Karlsruhe, Germany) with 10% fetal calf serum (FCS) (PAA Laboratories, Cölbe, Germany). Approximately 1.5×10^5 cells were seeded on a 35-mm cell culture plate and, 1 day later, co-transfected with 1-μg *pSEAP* and 1-μg *pcDNA3.1lacZ* (Invitrogen) using FuGene6 reagent (Roche Diagnostics, Germany). Cells were harvested on day 4 after transfection and the supernatants were stored at −20°C. Cells were washed twice with ice-cold PBS and stored at −80 °C. Alkaline phosphatase activity in the supernatant was measured with the Great EscAPe SEAP Genetic Reporter System (Clontech). Chemiluminescence was measured as relative light units (RLU) in a luminometer. For the determination of normalized ß-galactosidase activity, the transfected cells were lysed on cell culture plates for 1 h with 400-μl buffer containing 20 mM sodium phosphate, 1 mM $MgCl_2$ and 0.2% Triton X-100 at pH 7.8. Enzyme activity was measured in cell lysate aliquots containing 50-μg protein by using the chemiluminescence ELISA ß-Gal Kit (Roche Diagnostics). Normalization for transfection efficiency was performed in all experiments on the basis of β-galactosidase activity, i.e., alkaline phosphatase RLU were divided by the normalized ß-galactosidase RLU. Transfections were done in triplicate and each experiment was reproduced at least three times. Vector *pSEAP-Control*, which contains the $SV40_e$ promoter and *SV40* enhancer, served as a reference for promoter strength, and *pSEAP-Basic* (no control elements) as negative control.

3. Results

3.1. The crp gene

In order to characterize the genomic organization of the *crp* gene, we have used several cloning strategies, i.e., genomic walking, screening of a genomic library, and genomic PCR. Fig. 1A shows the *crp* gene: it spans approximately 12 kbp of DNA and it is comprised of 10 exons and nine introns. All splice junctions follow the GT/AG rule [26,27]. The *crp* exons range in size from 81 to 579 bp. Exon 1 with only 172 bp contains the 5′ non-coding region (see below for mapping of the transcription initiation site) and the first 10 bp of the coding region. Exon 10 (579 bp) contains the whole 3′ non-coding region. The exons 2 to 9 are of the symmetric class 1-1 [28,29] and comprise the major part of the coding region. Exon 2 is 105 bp long and, hence, somewhat longer than exons 3 to 8, each of them exhibit the same uniform size of 93 bp. The 72 bp long exon 9 encodes a shortened repeat. Fig. 1B shows the amino acid sequence deduced from the exons. There are seven uniform tandem repeats with Cys-X-Cys and Cys-Cys segments extending from amino acid position 39. Repeats E3 to E8 contain 31 residues and the carboxy-terminal repeat E9 lacks the last four residues. The 35 residues encoded by exon 2 also contain two Cys-X-Cys segments and two Cys residues at conserved positions, but not the Cys-Cys segment (Fig. 1B).

Fragments of known repetitive elements were identified in the phase 1 introns using sequence data from Rep-Base [30,31]. Introns range in size from 598 to 1359 bp, except for intron 6. The latter is about 2.3-kb long as determined by genomic restriction fragment analysis and only 628 bp



B

E1	M	A	S																														3
E2	H	A	H	A	E	A	H	**C**	G	**C**	G	T	E	**C**	H	**C**	***P***	K	G	S	**C**	G	S	G	**C**	G	T	S	T	T	A	G	S	**C**	G	38
E3					V	A	G	**C**	***P***	**C**	G	S	H	**C**	K	**C**	E	K	G	E	**C**	R	R	G	**C**	S	K	G	**C**	**C**	T	***P***	K	**C**	G	69
E4					V	E	G	**C**	***P***	**C**	G	S	Q	**C**	K	**C**	E	K	G	E	**C**	K	K	G	**C**	K	E	G	**C**	**C**	A	***P***	K	**C**	G	100
E5					V	A	G	**C**	S	**C**	G	S	G	**C**	K	**C**	E	K	G	E	**C**	K	***P***	G	**C**	T	K	R	**C**	**C**	G	T	K	**C**	G	131
E6					V	E	D	**C**	***P***	**C**	G	***P***	S	**C**	K	**C**	E	K	G	D	**C**	K	V	N	**C**	S	K	G	**C**	**C**	T	***P***	K	**C**	G	162
E7					V	V	G	**C**	***P***	**C**	G	S	Q	**C**	T	**C**	E	K	G	S	**C**	K	K	G	**C**	S	K	G	**C**	**C**	T	***P***	K	**C**	G	193
E8					M	Q	D	**C**	***P***	**C**	G	S	H	**C**	K	**C**	E	K	G	S	**C**	E	K	G	**C**	T	Q	G	**C**	**C**	A	***P***	K	**C**	D	224
E9					N	V	N	**C**	K	**C**	G	S	S	**C**	R	**C**	E	K	G	T	**C**	T	P	S	**C**	A	Q	T	**C**	**C**	A					251
cons	E2-9				V	X	G	**C**	P	**C**	G	S	X	**C**	K	**C**	E	K	G	E	**C**	K	K	G	**C**	S	K	G	**C**	**C**	T	P	K	**C**	G	

C

MRE-like repeat
5'-aatgtgtgttgatgtgttgaacagtacaggactgcccttgtggttcacattgt-3'

repeat I9
5'-aatcaattaactctcggtgaattctattaattg-3'

repeat A
5'-ctgaatgatatgcgtgtaattgattaatttcaagtgttatgtagcacgtgaagaggaaaacggaagatagagatg
aatggagacgttatataccatgatggacctgcctgtaggctgtgca-3'

repeat B
5'-aaatattttgataaatgatttaactgaattgtttccgatgttttatgtatttag-3'

repeat C
5'-aaaataatatcaatgacttcatgtaattaatattaaatcatttacacctatttgatcaaagatatctgtaatttc
tttgaaatttaatttatttaaatagttatatttttttatacttattgaagaaagtcttttcgacatctgtagttgt
ttatttaattttgttaaacag-3'

repeat iD
5'-aaaagctggtaaagctaaggttaatatcagtttcctcactcgaaagcgatgtcagtgacggtcgagcccaggatt
tgaaccccgcacaatccgtccgcagtcgagtatcttactgacttcgctactggctccagcca-3'

repeat E
5'-tcagaaagggcattcggttggtgagcgcctaggctcca-3'

Fig. 1. Genomic organization of the *crp* gene. (A) The schematic gene structure indicates the 10 exons (*E*) as numbered boxes interrupted by introns (*I*). The open bar in front of *E1* shows the characterized 5′-flanking region. Restriction map and localization of cloned DNA fragments are presented above and the structural organization of the CRP protein is depicted below. (B) Amino acid sequences encoded by the different exons. Cysteine and proline residues are in bold letters; *cons*, consensus sequence of the repeats encoded by *E*2–9. (C) Repeats of the *crp* gene. *MR-like repeat* is a tandem repeat in *I*7 with MRE-like heptanucleotide (underlined). *Repeat I9* is also a tandem repeat in *I*9. *Repeat A* occurs in *I*1 and *I*7; *repeat B* is located in *I*1 and *I*2; *repeat C* occurs in *I*2 and *I*4; *repeat iD* is an inverted repeat flanking *E*6; *repeat E* is localized in *I*8 and *I*9. Details are explained in Results.

downstream to exon 6 could be sequenced. A sequence stretch of intron 1 (position 567–862) shows 62% sequence identity to the L2A consensus sequence of the long interspersed nuclear element (LINE)-family [32]. Intron 2 contains overlapping sequence stretches with identities to mammalian transposon ZAPHOD (position 1971–2139, 66% identity) to *Dictyostelium* LTR-(long terminal repeat) retrotransposon Skipper (position 2086–2606, 67% identity), and to human transposon Mer3 (position 2097–2134, 89% identity) [33,34]. Intron 8 contains a sequence stretch (position 1501–1871) with 73% identity to *Drosophila* LTR-retrotransposon HMSBeagle [35]. In addition, there are 10 highly sequence-conserved unknown tandem repeats located in intron 7 (Fig. 1C), each containing the heptanucleotide *TGCCCTT* which resembles an MRE, but deviates in two positions from the MRE core consensus *TGCRCNC*

[36]. The tandem structure consists of nine repeats of 53 or 52 bp, one is 74 bp in length. Interspersed within the structure is a truncated 28-bp repeat. The 3′ repeat of the tandem extends into exon 8 with 29 bp. It is conspicuous, that putative MREs and HSEs (heat shock elements) [37,38] also accumulate in introns 1 (three MREs, three HSEs) and 8 (two MREs, three HSEs). Furthermore, a tandem repeat structure consisting of seven nearly identical 33-bp repeats is found in intron 9 (Fig. 1C). Remarkably, exon 6 is flanked by 140-bp inverted repeats with 85% identity (repeat iD) and direct repeats are located in introns 1 and 7 (121-bp repeat A, 99% identity), introns 1 and 2 (54-bp repeat B, 81% identity), introns 2 and 4 (174-bp repeat C, 85% identity), introns 8 and 9 (38-bp repeat E, 89% identity) (Fig. 1C). In intron 2, ZAPHOD overlaps with repeat B and Skipper overlaps partially with repeat C.

Three MT genes of the oligochaete *Lumbricus rubellus* encoding the isoforms mt2A, mt2B and mt2C were retrieved from EMBL data base (GenBank accession numbers: AJ299434, AJ299435, AJ299436) and compared to *crp* exons. Deduced coding regions of the three MT genes revealed that putative exons 3 and 4 not only have a comparable length of 99 and 86 bp, respectively, to *crp* exons but also show DNA sequence identities to *crp* exons between 54% and 71%. Comparison of deduced amino acid sequences of *crp* exons and *L. rubellus* MT exons 3 and 4 revealed identities up to 53% and similarities up to 77% (Fig. 2). There is also a remarkable high conservation of the Cys segment pattern among the lumbricid MTs and enchytraeid CRP exons.

3.2. The 5′-flanking region

Genomic walking resulted in the amplification of a 944-bp fragment of the 5′ flanking region from a *Bsh*1236I restricted DNA pool (Fig. 3A). The 3′-end of this PCR product is defined by the CRP2 primer, which includes the first four codons of the *crp* gene. The transcription initiation site(s) of the gene was mapped by RNase protection assay to be located at the thymine residue (+1) 163 bp upstream of the ATG codon. A minor initiation site is detected two residues upstream of the assigned +1 position (Fig. 3B).

Potential binding sites for transcription factors in the 770 bp 5′-flanking region were identified with the TRANSFAC database (Fig. 3A) [37]. The *crp* promoter is very rich in A+T residues (70%) and contains two nested putative TATA elements at position −26 and −28, respectively. There is no conformity of the major transcription initiation site with consensus initiator (Inr) sequence, only the Inr-conserved nucleotides A (+1) and T (+3) are found at the minor site [39]. There are in particular four putative MREs at proximal positions (−81, −122, −141, −192) with MREb and c in reverse orientation (Fig. 3A and C). MREb, MREc and MREd match the consensus heptanucleotide core, whereas MREa differs from the consensus in one position [36]. The proximal promoter (−170/−240) further contains binding sites for factors AP-1, AP-2, and AP-3 [40–42]. Among several potential GATA elements, there is at least one proximal located element at position −217 that fully coincides with the consensus binding site (reviewed in Ref. [43]). Binding sites for the basal transcription factors Sp1 (GC box) and CAAT binding protein are detectable at distal *crp* promoter positions (−682 and −476, respectively) [44–46]. Moreover, cAMP responsive elements (CREs) [47] are found at distal positions (−381, −510, −536) as well as xenobiotic response elements [48] (XREs, −396, −571) and an OctA1 binding site at −604 [49].

3.3. Activity of the crp promoter in HepG2 cells

Analysis of the *crp* promoter can be done only in a heterologous expression system, since there are not yet available any cell lines established from *E. buchholzi.* The human hepatoblastoma cell line HepG2 was chosen for transient transfections because these cells are known to strongly express endogenous and transfected MTs [50–53]. All constructs were linked to the SEAP reporter system. A *crp* promoter construct (−683/−8) containing all potential cis-acting motifs was first compared with the mouse *MT-II* promoter and the *SV40* early promoter (Fig. 4A). We have cloned a 410-bp genomic PCR fragment containing the

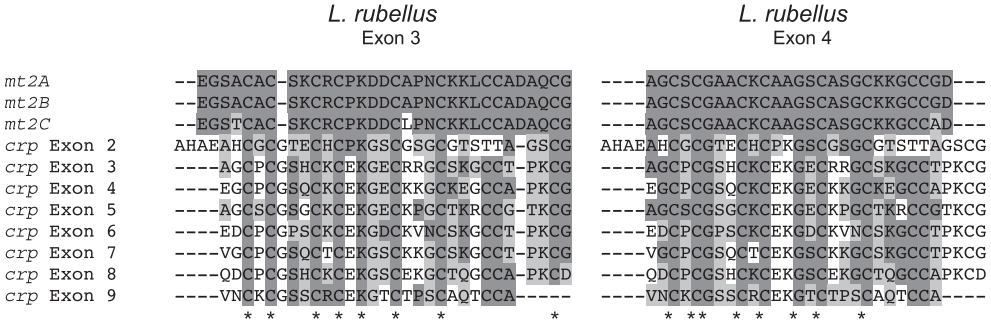


Fig. 2. Protein similarity of *L. rubellus* MT exons 3 and 4 to CRP repeats. Translated exons 3 and 4 of the *L. rubellus* mt2A, mt2B, mt2C genes (GenBank accession numbers: AJ299434, AJ299435, AJ299436) were compared to translated *crp* exons using *ClustalW* [68]. Identical residues are shaded dark grey, related residues are grey. Stars indicate residues conserved in all exons.

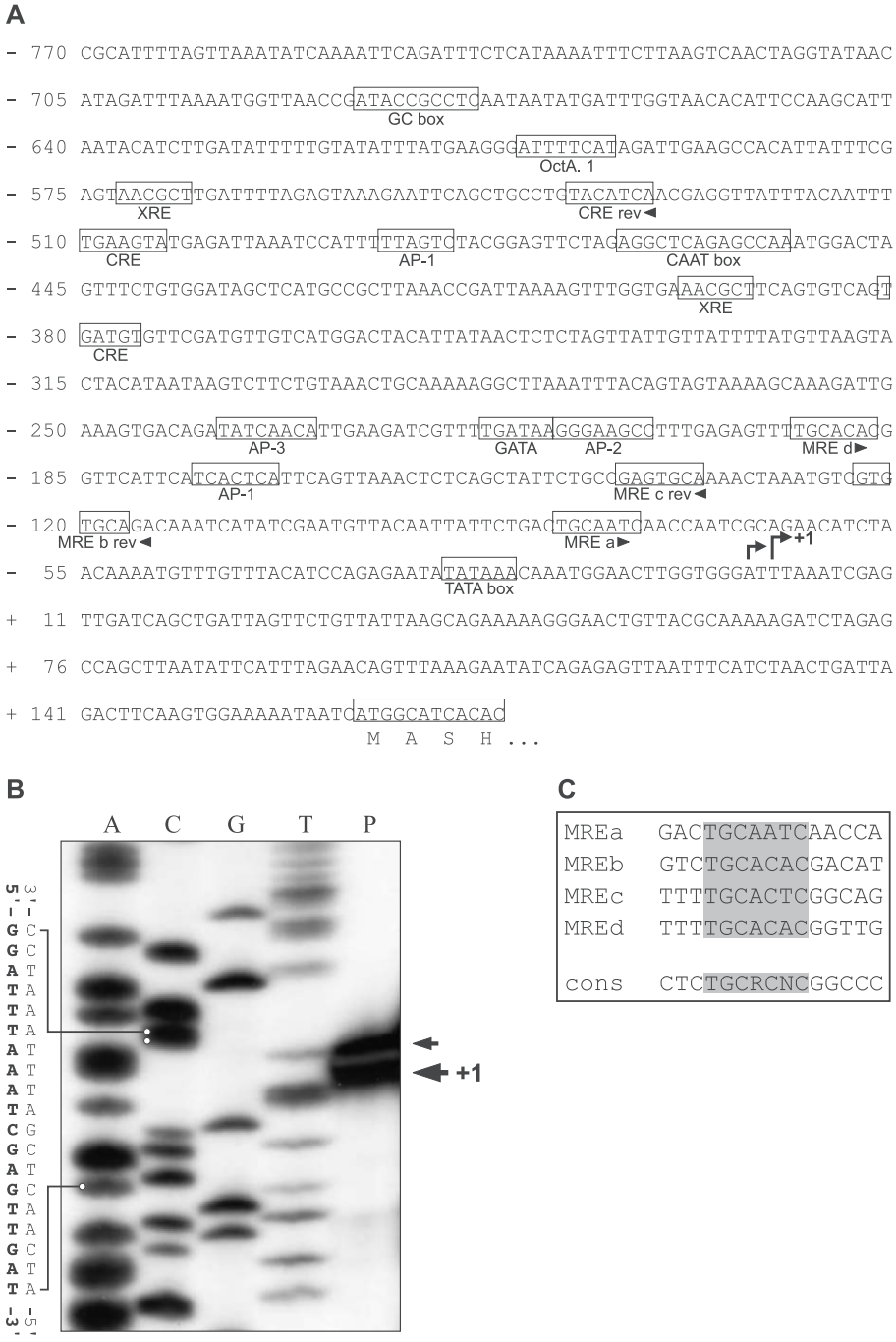


Fig. 3. The *crp* promoter. A 944-bp fragment of the 5′-flanking region was obtained from genomic walking using a *Bsh*1236I restricted uncloned DNA pool as described in "Materials and methods". (A) Putative regulatory *cis*-acting binding sites for transcription factors are marked by open boxes and arrowheads denote their orientation. The major transcription initiation site is denoted by +1, the identified minor start site is denoted by an arrow. The fragment also comprises the first four codons (open box). Numbers on the left indicate sequence positions relative to +1. (B) Mapping of the transcription initiation sites by a ribonuclease protection assay. The 944-bp promoter fragment was used as template to synthesize a radiolabeled antisense in vitro transcript, which was hybridized with poly $(A)^+$-RNA isolated from Cd-treated worms. Protected fragments were analysed on a sequence gel (arrowheads, lane *P*). A sequence reaction of the anti-sense strand of the same 944-bp promoter fragment was used as a size marker (lanes A, C, G, T). (C) Sequences of the four MREs (bold) localized in the *crp* promoter. Cons: MRE-consensus heptanucleotide core (bold) and flanking semi-conserved nucleotides derived from higher eukaryotic MT genes [36].

mouse MT-II (mMT-II) gene promoter with six MREs [36,54] in *pSEAP-Enhancer* plasmid (construct pMT-II) just as the *crp* promoter. The *crp* 5′ flanking region (construct −683/−8) promoted high levels of normalized SEAP expression in HepG2 cells and was set as 100% reference activity (Fig. 4A). In contrast, the mMT-II promoter is remarkably less active with only 13% activity. Also, the $SV40_e$ promoter only reached an activity level of 32%.

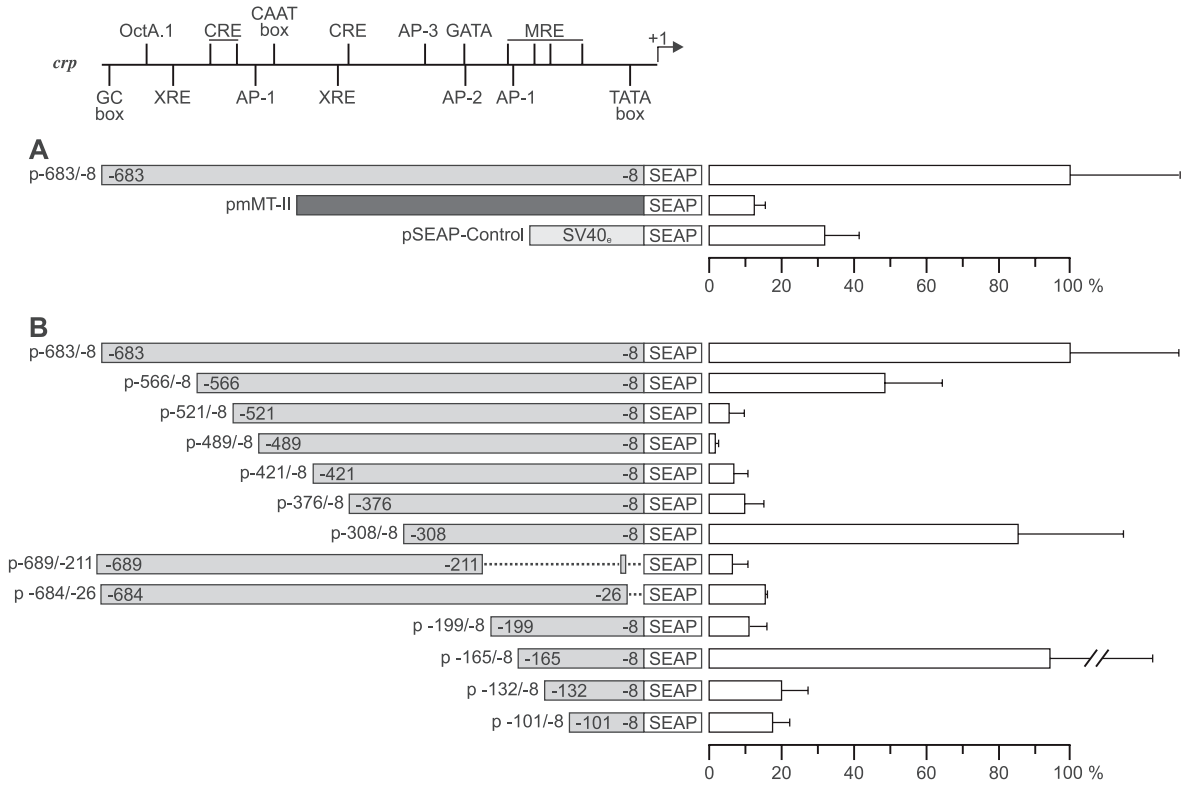


Fig. 4. Transcriptional activity of the *crp* 5′-flanking region. Fragments of the *crp* 5′-flanking region and a 410-bp fragment of the mouse MT-II promoter were generated by PCR and cloned in Xho*I* and *Hin*dIII restriction sites of vector *pSEAP-Enhancer*. Constructs were transfected into HepG2 cells. The ß-galactosidase expression vector *pcDNA3.1lacZ* was cotransfected as a control for transfection efficiency. SEAP activity was divided by the normalized ß-galactosidase activity to correct for transfection variability. Each experiment was done in triplicate and included transfection of *pSEAP-Basic* (no control elements). The data shown are mean values±standard deviation of at least three independent experiments. The activity of full *crp* promoter construct *p-683/-8* was set as 100% reference activity. (A) The construct *p-683/-8* was compared with mouse MT-II promoter (*pmMT-II*) and *SV40*$_e$ promoter (*pSEAP-Control*). (B) Activity of several deletion constructs of the *crp* promoter. Numbers indicate sequence positions relative to +1. Dotted lines indicate deleted sequences.

A series of promoter deletions fused to SEAP-reporter gene were constructed for the mapping of regulatory regions. The truncations revealed that the *crp* promoter could be subdivided in three regions of different influence on reporter gene activity (Fig. 4B). The distal promoter region −683/−521, which contained the GC box, the OctA.1 motif, one XRE, and one CRE, was very important for promoter activity, since only 48% of the SEAP activity of the full promoter was retained with the −566/−8 construct and a more extensive truncation to −521 reduced SEAP activity to very low levels of 5%. Three consecutive deletions of the central region from −521 to −376 also resulted in very low SEAP levels (<10%). Several putative motifs are localized within that region (Fig. 3A). However, reconstitution of about 85% of the full promoter activity was observed by deleting the −376/−309 region (construct −308/−8). This suggested that negative regulatory element(s) were possibly located in the central region from −521 to −309. However, cooperation between central region and the proximal *crp* promoter (−308/−8) is possibly required for the observed silencing effect.

Moreover, we tested the mutual influences of the distal and central promoter regions by generating the construct −689/−211, to which the TATA box region of *crp* was fused (Fig. 4B). The extremely decreased activity of 6.4% indicated that (i) the missing proximal region, containing the four MREs and the AP-2, AP-1 motifs, was essential for promoter function, and (ii) negative regulatory sequences of the central region possibly surpassed potential activating influences of distally located motifs.

The influence of the TATA box was investigated with the −683/−26 construct, which contained the full promoter without the TATA box region (Fig. 4B). Promoter activity was reduced by about 85%. Hence, it follows that the putative TATA box is an essential element for transcription initiation of the *crp* gene and that high *crp* promoter activity was only achieved by interaction of the TATA box with proximal located elements.

3.4. The importance of MREs for basal expression in HepG2 cells

The proximal construct −308/−8 with the TATA box, four MREs, AP-1, AP-2, and AP-3 elements achieved 85% of the SEAP activity of the full promoter construct −683/−8 (4B). Deletion of the AP-2 and AP-3 containing region (−308/−200) was devastating to the activity level of construct −199/−8 (11%). Though the proximal −199/−8

promoter with the four MREs was surprisingly inactive, a more extensive deletion of the AP-1 motif and MREd reconstituted a very high activity of construct −165/−8 (Fig. 4B). However, SEAP levels of that construct exhibited a higher variability in HepG2 than other constructs. Deletion of MREc revealed that the high activity of −165/−8 is dependent on the occurrence of this element (Fig. 4B). MREc was obviously crucial for high *crp* promoter activity in HepG2 cells. MREb did not participate in activation since the constructs −132/−8 (MREb and MREa) and −101/−8 (MREa) revealed comparably low level SEAP activities. The contribution of MREa to the 18% activity level of construct −101/−8 was not determined by an MREa-less construct. However, the −101/−8 region (TATA box region and MREa) displayed an activity level (18%) about threefold higher than the level promoted by construct −683/−211 to which the *crp* TATA box region was fused.

4. Discussion

The terrestric oligochaete worm *Enchytraeus* is able to survive in Cd-polluted environments, due to its 25-kDa CRP protein imparting Cd-tolerance [17–19]. Here, we have revealed a unique structure of the *crp* gene: It consists of 10 exons, in part repetitively arranged over a length of approximately 12 kb. The first exon, containing the 5′ non-coding region, encodes only the first three amino acids of the CRP protein, exon 2 the next 35 amino acids, exons 3 to 8 six tandemly arranged repeats of the same size of 31 amino acids, and exon 9 a truncated repeat of 27 amino acids, respectively. There has not yet been identified any gene encoding a protein such as the CRP. Only the MTs exhibit similarities to CRP in so far as both proteins reveal a conspicuous accumulation of Cys-Cys and Cys-X-Cys motifs. The Cys-rich metal-binding protein CRP should be therefore defined as a MT. However, the "classical" MT genes are much smaller than the *crp* gene. Vertebrate MT genes display a 3 exon/2 intron structure, whereas the invertebrate MT genes are organized more heterogeneously [10]. For instance, the MT gene of the protozoan *Tetrahymena* has no intron at all; the MT genes of the nematode worm *C. elegans* and the insect *Drosophila* exhibit a 2 exon/1 intron structure [11,55], and the terrestric earthworm *L. rubellus* has three MT genes with at least 4 exons/3 introns (GenBank accession numbers: AJ299434, AJ299435, AJ299436).

A highly remarkable peculiarity of the *crp* gene, which is not present in the numerous MT genes identified to date, is the fact that six exons reveal an identical size and encode highly identical protein sequences. This indicates that one exon encodes one functional domain, and, additionally, it appears as if these domain repeats are evolutionary originated by exon duplications [56–59]. This view is supported by the following facts. First, the symmetric class 1-1 exons 2 to 9 are separated by phase 1 introns. Second, Exon 6 is flanked at both ends by inverted repeats suggestive of transposable processes. Third, there are a series of other known transposon-like and retroviral-like/nonretroviral retrotransposon-like sequences in the different introns suggesting transposon- or retrotransposon-driven gene rearrangements of the ancestral *crp* gene during evolution [60,61]. The process of exon shuffling by LINE-1s retrotransposition was recently demonstrated in cultured human cells [62]. The *crp* gene and the diverse MT genes, despite the dissimilarities in their actual structural organization [63], may have evolved from a common smaller ancestral gene. This view is corroborated by DNA and protein sequence comparisons between MTs and CRP repeats: The putative exons 3 and 4 of mt2A, mt2B, and mt2C of the oligochaete *L. rubellus* do not only have comparable sizes to *crp* exons, but also exhibit high sequence similarity. Moreover, CRP repeats and different mouse and human MTs exhibit protein identities ranging between 42% and 53% [20].

A previous study has shown that the *crp* gene is not constitutively expressed in *E. buchholzi* [19]. However, in HepG2 cells, the *crp* promoter reveals a basal expression of a significantly higher level than the mMT-II promoter, which contains six MREs [36]. The proximal promoter region −308/−8, containing AP-1, AP-2, AP-3 as well as the four MREs, is sufficient for strong basal *crp* promoter activity but highest activity levels resulted only from the complete 5′ flanking region including the GC box at −682. The transcription factor Sp1, which binds to the GC box, as well as activator proteins AP-1 and AP-2 regulate basal activity of MT genes [40,41,64,65]. The TATA box and the MRE-containing region have a crucial impact on basal *crp* promoter activity in HepG2 cells. Proximal and distal segments of the *crp* promoter do not compensate for the loss of the TATA element. The relevance of the TATA box for basal expression has been demonstrated for human MT-I gene promoters [53,54]. MREs are implicated in basal regulation of MT genes, as shown for rainbow trout MT-B gene [66], human MT-IG [53] and human MT-IIA [67]. Interestingly, MREa of the trout MT-B gene promoter contributes to basal expression both in fish cells and in mammalian cells [66]. MREa of the human MT-IG promoter is obviously crucial for the basal function of the TATA box and for efficient metal-inducibility of the upstream MREs [53]. The four proximal MREs of *crp* promoter contribute differently to basal activity in HepG2 cells and deletion of the whole MRE region is devastating to promoter activity. In particular, MREc is essential for basal promoter function by preserving a very strong expression.

References

[1] J.O. Nriagu, J.M. Pacyna, Quantitative assessment of worldwide contamination of air, water and soils by trace metals, Nature 333 (1988) 134–139.

[2] M. Stoeppler, Cadmium, in: E. Merian (Ed.), Metals and Their Compounds in the Environment, VCH, Weinheim, 1991, pp. 803–851.
[3] B.L. Vallee, D.D. Ulmer, Biochemical effects of mercury, cadmium, and lead, Ann. Rev. Biochem. 41 (1972) 91–128.
[4] K.B. Jacobson, J.E. Turner, The interaction of cadmium and certain other metal ions with proteins and nucleic acids, Toxicology 16 (1980) 1–37.
[5] J.H.R. Kägi, Evolution, structure, and chemical activity of class I metallothioneins: an Overview, in: K.Y. Suzuki, N. Imura, M. Kimura (Eds.), Metallothioneins III, Birkhäuser Verlag, Basel, 1993, pp. 29–56.
[6] C.D. Klaassen, J. Liu, S. Choudhuri, Metallothionein: an intracellular protein to protect against cadmium toxicity, Annu. Rev. Pharmacol. Toxicol. 39 (1999) 267–294.
[7] A.T. Miles, G.M. Hawksworth, J.H. Beattie, V. Rodilla, Induction, regulation, degradation, and biological significance of mammalian metallothioneins, Crit. Rev. Biochem. Mol. Biol. 35 (2000) 35–70.
[8] P. Coyle, J.C. Philcox, L.C. Carey, A.M. Rofe, Metallothionein: the multipurpose protein, Cell. Mol. Life Sci. 59 (2000) 627–647.
[9] J.H.R. Kägi, Overview of metallothionein, Methods Enzymol. 205 (1979) 613–626.
[10] L.A. Samson, L. Gedamu, Molecular analyses of metallothionein gene regulation, Prog. Nucleic Acid Res. Mol. Biol. 59 (1998) 257–288.
[11] M.K. Sakharkar, P. Kanguene, T.W. Woon, T.W. Tan, P.R. Klatkar, M. Long, S.J. de Souza, IE-Kb: intron exon knowledge base, Bioinformatics 16 (2000) 1151–1152.
[12] R.D. Palmiter, Molecular biology of metallothionein expression, Experientia, Suppl. 52 (1987) 63–80.
[13] G.W. Stuart, P.F. Searle, H.Y. Chen, R.L. Brinster, R.D. Palmiter, A 12-base pair DNA motif that is repeated several times in metallothionein gene promoters confers metal regulation to a heterologous gene, Proc. Natl. Acad. Sci. U. S. A. 81 (1984) 7318–7322.
[14] G.W. Stuart, P.F. Searle, R.D. Palmiter, Identification of multiple metal regulatory elements in mouse metallothionein-I promoter by assaying synthetic sequences, Nature 317 (1985) 828–831.
[15] V.C. Culotta, D.H. Hamer, Fine mapping of a mouse metallothionein gene metal response element, Mol. Cell. Biol. 9 (1989) 1376–1380.
[16] H. Stone, J. Overnell, Non-metallothionein cadmium-binding proteins, Comp. Biochem. Physiol. C, Comp. Pharmacol. Toxicol. 80 (1985) 9–14.
[17] J. Willuhn, H.-P. Schmitt-Wrede, H. Greven, F. Wunderlich, Cadmium-induced mRNA encoding a non-metallothionein 33 kDa protein in *Enchytraeus buchholzi*, Ecotoxicol. Environ. Saf. 29 (1994) 93–100.
[18] J. Willuhn, H.-P. Schmitt-Wrede, H. Greven, F. Wunderlich, cDNA cloning of a cadmium-inducible mRNA encoding a novel cysteine-rich, non-metallothionein 25-kDa protein in an enchytraeid earthworm, J. Biol. Chem. 269 (1994) 24688–24691.
[19] J. Willuhn, A. Otto, H-.P. Schmitt-Wrede, F. Wunderlich, Earthworm gene as indicator of bioefficacious cadmium, Biochem. Biophys. Res. Commun. 220 (1996) 581–585.
[20] S. Tschuschke, H.-P. Schmitt-Wrede, H. Greven, F. Wunderlich, Cadmium resistance conferred to yeast by a non-metallothionein-encoding gene of the earthworm *Enchytraeus*, J. Biol. Chem. 277 (2002) 5120–5125.
[21] M.L. Chernaik, P.C. Huang, Differential effect of cysteine-to-serine substitutions in metallothionein on cadmium resistance, Proc. Natl. Acad. Sci. U. S. A. 88 (1991) 3024–3028.
[22] J.H.R. Kägi, A. Schäffer, Biochemistry of metallothionein, Biochemistry 27 (1988) 8509–8515.
[23] J.H.R. Kägi, Y. Kojima, Chemistry and biochemistry of metallothionein, Experientia, Suppl. 52 (1987) 25–61.
[24] P.D. Siebert, A. Chenchik, D.E. Kellog, K.A. Lukyanov, S.A. Lukyanov, An improved PCR method for walking in uncloned DNA, Nucleic Acids Res. 23 (1995) 1087–1088.
[25] J. Krücken, O. Stamm, H.P. Schmitt-Wrede, A. Mincheva, P. Lichter, F. Wunderlich, Spleen-specific expression of the malaria-inducible intronless mouse gene imap38, J. Biol. Chem. 274 (1999) 24383–24391.
[26] R.M. Stephens, T.D. Schneider, Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites, J. Mol. Biol. 228 (1992) 1124–1136.
[27] S.M. Mount, C. Burks, G. Hertz, G.D. Stormo, O. White, C. Fields, Splicing signal in *Drosophila*: Intron size, information content, and consensus sequences, Nucleic Acids Res. 20 (1992) 4255–4262.
[28] E.V. Kriventseva, M.S. Gelfand, Statistical analysis of the exon-intron structure of higher and lower eukaryote genes, J. Biomol. Struct. Dyn. 17 (1999) 281–288.
[29] M. Tomita, N. Shimizu, D.L. Brutlag, Introns and reading frames: correlation between splicing sites and their codon positions, Mol. Biol. Evol. 13 (1996) 1219–1223.
[30] J. Jurka, Repeats in genomic DNA: mining and meaning, Curr. Opin. Struct. Biol. 8 (1998) 333–337.
[31] J. Jurka, Repbase update: a database and an electronic journal of repetitive elements, Trends Genet. 9 (2000) 418–420.
[32] A.F.A. Smit, A.D. Riggs, MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation, Nucleic Acids Res. 23 (1995) 98–102.
[33] J. Jurka, Novel families of interspersed repetitive elements from the human genome, Nucleic Acids Res. 18 (1990) 137–141.
[34] P. Leng, D.H. Klatte, G. Schumann, J.D. Boeke, T.L. Steck, Skipper, an LTR retrotransposon of Dictyostelium, Nucleic Acids Res. 26 (1998) 2008–2015.
[35] M.P. Snyder, D. Kimbrell, M. Hunkapiller, R. Hill, J. Fristom, N.A. Davidson, A transposable element that splits the promoter region inactivates a *Drosophila* cuticle protein gene, Proc. Natl. Acad. Sci. U. S. A. 79 (1982) 7430–7434.
[36] J. Imbert, V. Culotta, P. Fürst, L. Gedamu, D. Hamer, Regulation of metallothionein gene transcription by metals, Adv. Inorg. Biochem. 8 (1990) 139–164.
[37] K. Quandt, K. Frech, H. Karas, E. Wingender, T. Werner, MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data, Nucleic Acids Res. 23 (1995) 4878–4884.
[38] J.J. Bonner, C. Ballou, C.L. Fackenthal, Interactions between DNA-bound trimers of the yeast heat shock, Mol. Cell. Biol. 14 (1994) 501–508.
[39] S.T. Smale, Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes, Biochim. Biophys. Acta 1351 (1997) 73–88.
[40] W. Lee, A. Haslinger, M. Karin, R. Tjian, Activation of transcription by two factors that bind promoter and enhancer sequences of the human metallothionein gene and SV40, Nature 325 (1987) 368–372.
[41] M. Imagawa, R. Chiu, M. Karin, Transcription factor AP-2 mediates induction by two different signal-transduction pathways: protein kinase C and cAMP, Cell 51 (1987) 251–260.
[42] F. Mercurio, M. Karin, Transcription factors AP-3 and AP-2 interact with the SV40 enhancer in a mutually exclusive manner, EMBO J. 8 (1989) 1455–1460.
[43] R.K. Patient, J.D. McGhee, The GATA family (vertebrates and invertebrates), Curr. Opin. Genet. Dev. 12 (2002) 416–422.
[44] K.A. Jones, R. Tjian, Sp1 binds to promoter sequences and activates herpes simplex virus 'immediate-early' gene transcription in vitro (1985), Nature 317 (1985) 179–182.
[45] M. Bellorini, K. Zemzoumi, A. Farina, J. Berthelsen, G. Piaggio, R. Mantovani, Cloning and expression of human NF-YC, Gene 193 (1997) 119–125.
[46] R. Mantovani, A survey of 178 NF-Y binding CCAAT boxes, Nucleic Acids Res. 26 (1998) 1135–1142.
[47] S.M. Crosson, G.F. Davies, W.J. Roesler, Cyclic AMP-stimulated accumulation of the cAMP response element binding protein can occur without changes in gene expression, Biochem. Biophys. Res. Commun. 227 (1996) 915–920.

[48] N. Matsushita, K. Sogawa, M. Ema, A. Yoshida, Y. Fujii-Kurijama, A factor binding to the xenobiotic responsive element (XRE) of P-4501A1 gene consists of at least two helix-loop-helix proteins, Ah receptor and Arnt, J. Biol. Chem. 268 (1993) 21002–21006.
[49] R. Sturm, T. Baumruker, B.R. Franza Jr., W. Herr, A 100-kD HeLa cell octamer binding protein (OBP100) interacts differently with two separate octamer-related sequences within the SV40 enhancer, Genes Dev. 1 (1987) 1147–1160.
[50] C. Sadhu, L. Gedamu, Regulation of human metallothionein (MT) genes. Differential expression of MTI-F, MTI-G, and MTII-A genes in the hepatoblastoma cell line (HepG2), J. Biol. Chem. 263 (1988) 2679–2684.
[51] R. Foster, L. Gedamu, Functional analyses of promoter elements responsible for the differential expression of the human metallothionein (MT)-IG and MT-IF genes, J. Biol. Chem. 266 (1991) 9866–9875.
[52] N.-W. Shworak, T. O'Connor, N.C.W. Wong, L. Gedamu, Distinct TATA motifs regulate differential expression of human metallothionein I genes MT-IF and MT-IG, J. Biol. Chem. 268 (1993) 24460–24466.
[53] S.L.-A. Samson, W.P. Paramchuk, N.W. Shworak, L. Gedamu, Functional analyses of the human metallothionein-IG gene. In vitro and in vivo studies, J. Biol. Chem. 270 (1995) 25194–25199.
[54] P.F. Searle, B.L. Davison, G.W. Stuart, T.M. Wilkie, G. Norstedt, R.D. Palmiter, Regulation, linkage, and sequence of mouse metallothionein I and II genes, Mol. Cell. Biol. 4 (1984) 1221–1230.
[55] J.H. Freedman, L.W. Slice, D. Dixon, A. Fire, C.S. Rubin, The novel metallothionein genes of *Caenorhabditis elegans*, J. Biol. Chem. 268 (1993) 2554–2564.
[56] L. Patthy, Genome evolution and the evolution of exon-shuffling—a review, Gene 238 (1999) 103–114.
[57] J.A. Kolkman, W.P. Stemmer, Directed evolution of proteins by exon shuffling, Nat. Biotechnol. 19 (2001) 423–428.
[58] W.-H. Li, Z. Gu, H. Wang, A. Nekrutenko, Evolutionary analyses of the human genome, Nature 409 (2001) 847–849.
[59] M. Long, Evolution of novel genes, Curr. Opin. Genet. Dev. 11 (2001) 673–680.
[60] H.S. Malik, W.D. Burke, T.H. Eickbush, The age and evolution of non-LTR retrotransposable elements, Mol. Biol. Evol. 16 (1999) 793–805.
[61] A.F.A. Smit, Interspersed repeats and other mementos of transposable elements in mammalian genomes, Curr. Opin. Genet. Dev. 9 (1999) 657–663.
[62] J.V. Moran, R.J. DeBerardinis, H.H. Kazazian Jr., Exon shuffling by L1 retrotransposition, Science 283 (1999) 1530–1534.
[63] P.A. Binz, J.H.R. Kägi, Metallothionein: molecular evolution and classification, in: C.D. Klaassen (Ed.), *Metallothionein IV*, BirkhäuserVerlag, Basel, 1999, pp. 7–13.
[64] M. Karin, A. Haslinger, H. Holtgreve, R.I. Richards, P. Krauter, H.M. Westphal, M. Beato, Characterization of DNA sequences through which cadmium and glucocorticoid hormones induce human metallothionein-IIA gene, Nature 308 (1984) 513–519.
[65] E. Saint-Jacques, M.-J. April, C. Séguin, Structure and metal-regulated expression of the gene encoding *Xenopus laevis* metallothionein-A, Gene 160 (1995) 201–206.
[66] S.L. Samson, L. Gedamu, Metal-responsive elements of the rainbow trout metallothionein-B gene function for basal and metal-induced activity, J. Biol. Chem. 270 (1995) 6864–6871.
[67] S. Koizumi, K. Suzuki, Y. Ogra, H. Yamada, F. Otsuko, Transcriptional activity and regulatory protein binding of metal-responsive elements of the human metallothionein-IIA gene, Eur. J. Biochem. 259 (1999) 635–642.
[68] J.D. Thompson, D.G. Higgins, T.J. Gibson, ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap-penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.