A complex pattern of evolutionary conservation and alternative polyadenylation within the long 3'-untranslated region of the methyl-CpG-binding protein 2 gene (*MeCP2*) suggests a regulatory role in gene expression

Johannes F. Coy[1], Zdenek Sedlacek[1,3], Dietmar Bächner[2], Hajo Delius[1] and Annemarie Poustka[1,+]

[1]Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany, [2]Institut für Zellbiochemie und Klinische Neurobiologie, Universitätskrankenhaus Eppendorf, Martinistraße 52, 20246 Hamburg, Germany and [3]Institute of Biology and Medical Genetics, Second Medical School, Charles University, V uvalu 84, 15006 Prague, Czech Republic

A systematic search for expressed sequences in the human Xq28 region resulted in the isolation of 8.5 kb large contigs of human and murine cDNAs with no apparent conserved open reading frames. These cDNAs were found to be derived from the 3'-untranslated region (3'-UTR) of the methyl-CpG-binding protein 2 gene (*MeCP2*). This long 3'-UTR is part of an alternatively polyadenylated, 10.1 kb *MeCP2* transcript which is differentially expressed in human brain and other tissues. RNA *in situ* hybridization to sections of mouse embryo and adult tissues of an *Mecp2* 3'-UTR probe showed ubiquitous low level expression in early organogenesis and enhanced expression in the hippocampus during formation of the differentiated brain. Sequence comparison between the human and mouse homologues revealed several blocks of very high conservation separated by less conserved sequences. Additional support for a domain-like conservation pattern of the long 3'-UTR of the *MeCP2* gene was obtained by examining conservation in the chimpanzee, orangutan, macaque, hamster, rat and kangaroo. The minimum free energy distribution for the predicted RNA secondary structure was very similar in human and mouse sequences. In particular, the conserved blocks were predicted to be of high minimum free energy, which suggests weak secondary structure with respect to RNA folding. The fact that both the sequence and predicted secondary structure have been highly conserved during evolution suggests that both the primary sequence and the three-dimensional structure of the 3'-UTR may be important for its function in post-transcriptional regulation of MeCP2 expression.

INTRODUCTION

Methyl-CpG-binding protein 2 (MeCP2), an abundant chromosomal protein, was identified in 1992 (1) based on its selective binding to methylated CpG dinucleotides in mammalian genomic DNA. The methylation of the C residues in CpG dinucleotides was associated with transcriptionally repressed chromatin states and with transcriptional silencing of genes (2,3). Recently, a role for the MeCP2 protein in guiding a complex containing histone deacetylase to methylated chromatin domains was proposed. Histone deacetylation then induces the transcriptionally inactive condensed chromatin structure (4,5). Mouse Mecp2 is dispensable for the growth and *in vitro* differentiation of stem cells, but its absence in embryos causes severe developmental defects and early embryonic lethality (6), similar to defects due to absence of another key component of the methylation pathway, DNA methyltransferase (7). Little is known, however, about the regulation of this gene.

As a part of the functional analysis of the human Xq28 region, we have developed a method based on the hybridization of radioactively labelled complex cDNA probes derived from different pig and human tissues to cosmid clones gridded onto nylon filters and to blots of restriction digests of these clones (8). This approach led to the identification and isolation of several new genes (9,10). The characterization of another transcribed region identified by this method initially revealed a very unusual cDNA contig with no open reading frame (ORF), highly varying discontinuous conservation between human and mouse and no introns in its underlying genomic sequence. Precise mapping and extension of this cDNA contig has shown that it belongs to a >8.5 kb long 3'-untranslated region (3'-UTR) of the *MeCP2* gene. It became clear that alternative

+To whom correspondence should be addressed. Tel: +49 6221 424742; Fax: +49 6221 423454; Email: a.poustka@dkfz-heidelberg.de

polyadenylation of this human X-linked gene gives rise to one 1.8 kb transcript (11) or several much longer transcripts with lengths of up to 10.1 kb, differentially expressed in various tissues and containing different lengths of the long 3'-UTR. We also cloned the complete mouse 10.1 kb *Mecp2* transcript and compared it with the 10.1 kb human transcript. Besides the expected conservation of the ORFs between human and mouse, a very high conservation of several blocks of sequences in the long 3'-UTR sequences could be detected predicting a conserved three-dimensional RNA folding.

Differential expression of the long and short *MeCP2* mRNA variants seen in brain and other tissues points to a possible important role of this sequence in methylation-associated gene regulation. The 3'-UTRs of genes have been shown to act in post-transcriptional regulation of gene expression, such as the modulation of mRNA degradation and stability, nucleocytoplasmic transport, sorting and localization of mRNAs in the cytoplasm, modulation of translation and control of the coding capacity of transcripts (12). The isolation and analysis of the long variant of the *MeCP2* 3'-UTR provides the opportunity to study the possible role of this sequence in modulation of *MeCP2* protein expression.

RESULTS

Isolation of the long 3'-UTR variant of the human *MeCP2* gene

To identify transcribed sequences in human Xq28, a region-specific cosmid library (13) was screened with radioactively labelled complex cDNA isolated from pig brain and muscle (8). A cosmid Qc8D3 was detected, which showed a hybridization signal with these probes. A Southern blot of a restriction digest of this cosmid was then hybridized with the same probes. A positive 4.5 kb long *Hin*dIII restriction fragment was isolated and used to screen a region-specific cDNA library (14). This led to the isolation of one positive cDNA clone, 31g3B4. Sequencing of this 1 kb long cDNA clone and a database search revealed no significant homologies to known sequences.

To extend the sequence, clone 31g3B4 was hybridized to a conventional human fetal brain cDNA library and 15 positive cDNA clones were isolated and sequenced. One end of the sequence of the cDNA contig overlapped with the 3'-end of the human *MeCP2* cDNA sequence (GenBank accession no. X89430) (11). The hybridization probe from the 3'-end of the cDNA contig was generated by PCR and used to screen the same cDNA library. Another 25 new cDNA clones were identified. A new end probe was used to screen the cDNA library once more and 19 new cDNA clones could be isolated. Finally, an additional walk extended the novel cDNA contig to 8323 bp followed by a poly(A) tail indicating the 3'-end of the gene. Northern blot hybridization was performed with one of the overlapping cDNA clones (JFC484) and a transcript of ~10.1 kb was detected in all fetal and adult tissues tested (Fig. 1b and d).

In addition to the cDNA sequence of D'Esposito *et al.* (11), which joined the *MeCP2* gene with the new cDNA contig, an RT–PCR experiment on a mixture of RNAs using primers MEC2 and D44 confirmed that the *MeCP2* gene and the 8323 bp long cDNA contig belong to one transcript (data not shown). Furthermore, a probe (MEC1-MEC3) representing the
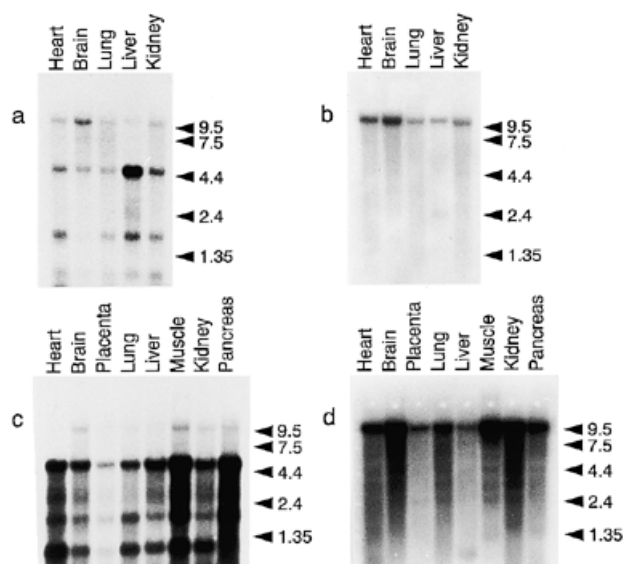

Figure 1. Expression pattern of the human *MeCP2* gene on northern blots of poly(A)⁺ mRNA from different human fetal (**a** and **b**) and adult (**c** and **d**) tissues analysed with a probe MEC1-MEC3 from the coding region (a and c) and with a probe JFC484 derived from the long variant of the 3'-UTR (b and d). The size markers are in kb.

coding region of the human *MeCP2* gene was generated by RT–PCR and hybridized to the northern blots, which had formerly been hybridized with the 3'-UTR probe JFC484. As expected, the large transcripts detected by the 3'-UTR probe are also detected by the coding region probe. The smaller transcripts (1.8 and 5 kb, respectively) of the *MeCP2* gene are only detected with the MEC1-MEC3 probe. The ratio between the 10.1 and 5 kb transcripts changes significantly during development of the brain. Whereas the 10.1 kb transcript is more abundant than the 5 kb transcript in the fetal brain, the 5 kb transcript is more abundant in the adult brain (Fig. 1a and c).

These results demonstrate that we have isolated the long 3'-UTR of the 10.1 kb long *MeCP2* transcript. Database comparisons reveal no significant sequence homology, except for the recently released 113 kb long genomic sequence of the *MeCP2* locus (GenBank accession no. AF030876) and many expressed sequence tag (EST) sequences. Most of the 3'-ESTs are identical to the 3'-end of the cDNA contig. A minority of the 3'-ESTs cluster ~2.9 kb upstream of the 3'-end of the contig and define another alternatively used poly(A) addition site (Fig. 2b), which can give rise to a 7.2 kb long *MeCP2* transcript. Transcription of the *MeCP2* gene is directed from the telomere to the centromere, on the complement of the published genomic sequence. The poly(A) addition site of the 1.8 kb *MeCP2* transcript is located at nucleotide position 21711. The new cDNA contig covers the complement of the genomic sequence up to base 13284 and both sequences are collinear with an absence of any introns. At the 3'-end, the 10.1 kb long variant of the *MeCP2* gene is flanked at a distance of <2 kb by the interleukin-1 receptor-associated kinase gene (*IRAK*) (15) (Fig. 2a).

Comparative sequence and expression analysis of human and mouse *MeCP2*

Because a comparison between human and murine homologues could identify regions that are conserved and therefore
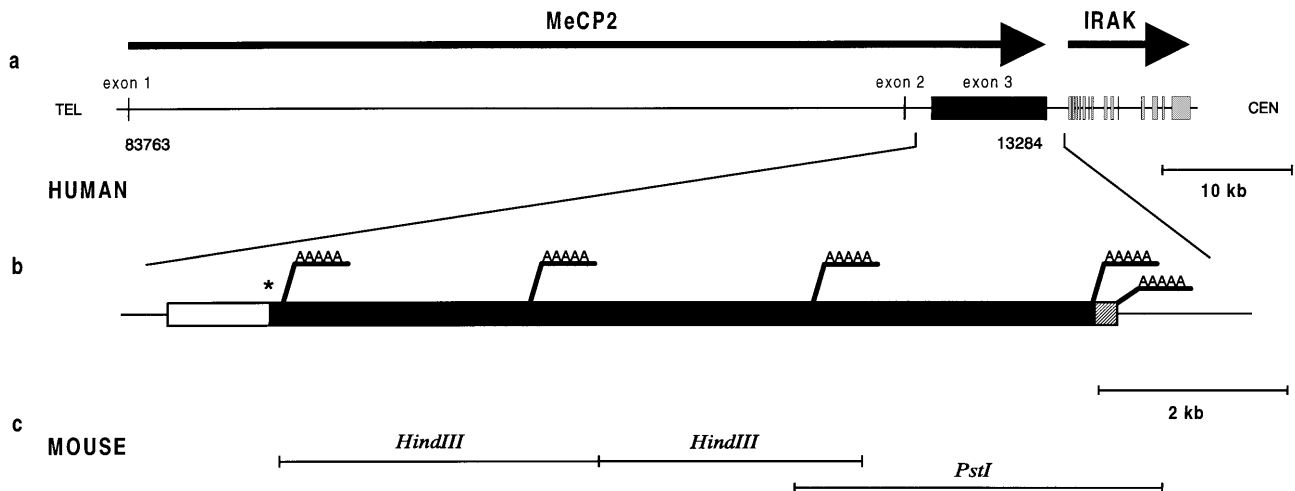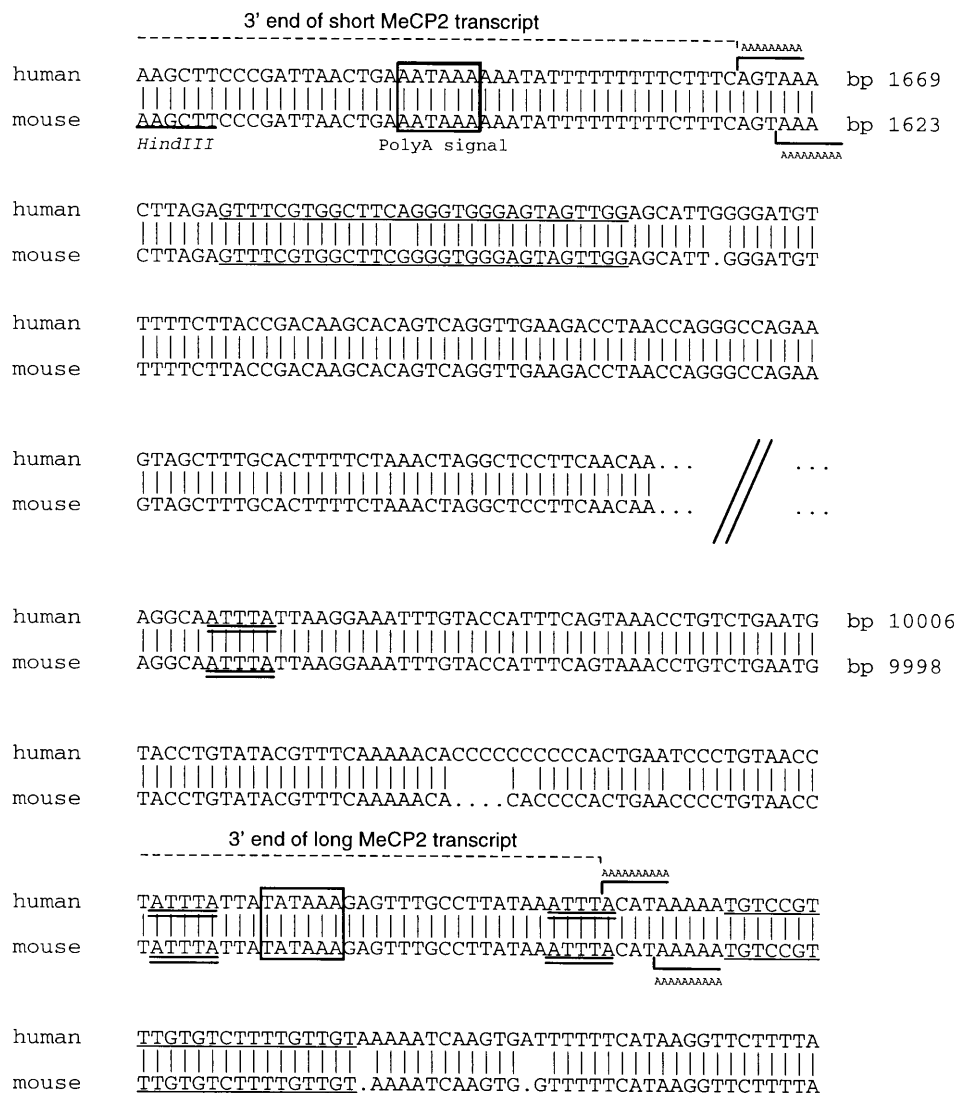
Figure 2. Location and structure of the human and murine *MeCP2* gene. (**a**) Physical map of the human genomic region showing the three exons of the *MeCP2* gene (black rectangles), 14 exons of the *IRAK* gene (grey rectangles), direction of transcription (arrows) and the distance between the two genes. Numbers correspond to the 5'- and 3'-ends of the published genomic sequence of the *MeCP2* gene (GenBank accession no. AF030876). (**b**) A detail of the third exon of the *MeCP2* gene. The positions of the different polyadenylation sites in the *MeCP2* gene are shown. The open rectangle represents the 3'-end of the ORF and the asterisk indicates the stop codon. The hatched box indicates the extension of the 3'-UTR by a rat EST (see text for details). (**c**) The genomic restriction fragments subcloned from the mouse P1 clone covering the 3'-end of the murine *Mecp2* gene.

possibly functionally important, we cloned and sequenced the complete transcript of the mouse *Mecp2* gene. To analyse the mouse genomic *Mecp2* locus, human cDNA clones JFC88 and JFC484 (from 5'- and 3'-ends of the UTR, respectively) were used to screen a mouse genomic P1 library (16) at reduced stringency. One positive clone, PM703EO165, was identified. This P1 clone was digested with *Pst*I and *Hin*dIII, subcloned into the Bluescript SKII vector and positive subclones were sequenced. The same human cDNA probes were also used to screen a mouse fetal brain cDNA library. Fourteen positive cDNA clones were identified and sequenced. Sequence analysis revealed collinearity between the mouse 8323 bp cDNA contig and genomic sequences. Two *Hin*dIII and one *Pst*I genomic restriction fragments cover the complete mouse *Mecp2* long 3'-UTR (Fig. 2c). To enable a comparison between the complete human and mouse *MeCP2* transcripts, the mouse short *Mecp2* transcript was also cloned by a PCR-based approach using primers MECP-CON-1 and MECP-MOU-4. A PCR reaction with an aliquot of a mouse fetal brain cDNA library with these primers resulted in amplification of a specific DNA fragment of 1284 bp, which was sequenced and showed high homology (90.7 and 95.5%, respectively) to the human and rat *MeCP2* sequences. Another PCR primer (MEC-MOU-11), located at the end of this sequence, was used in combination with primer MEC-MOU-18, located in the beginning of the 8323 bp mouse cDNA contig, to isolate the missing mouse cDNA sequence. PCR on the mouse fetal brain cDNA library and on total genomic mouse DNA resulted in the amplification of specific DNA fragments of the same size. Similar to human, this result demonstrated the absence of an intron between primers MEC-MOU-11 and MEC-MOU-18 and sequencing of both fragments revealed no sequence differences. This sequence represented the end of the mouse ORF and the 5'-end of the UTR. By combining the sequences of the 8323 bp cDNA contig and the sequences of the two mouse

PCR products, the complete sequence of the 10.1 kb long mouse *Mecp2* transcript could be assembled.

Figure 3 shows a comparison of the ends of the 1.8 and 10.1 kb long human and mouse 3'-UTRs of the *MeCP2* gene. Whereas a common AATAAA poly(A) signal is detectable in the 3'-end of the 1.8 kb *MeCP2* transcript, no canonical poly(A) signal is present in the 3'-end of the 10.1 kb long variants of the human and murine genes, although they contain a poly(A) tail. Instead, there is a sequence TATAAA in both human and mouse, which resembles the poly(A) signal and is polyadenylated with ~20% efficiency compared with the AATAAA sequence (17). No A-rich stretch is present in genomic DNA which could be responsible for artifactual internal priming of cDNAs. The third alternative poly(A) addition site defined by clustering of a minority of human 3'-ESTs has an AATAAA sequence. There are obvious GT elements (18) immediately downstream of the poly(A) addition sites of the 1.8, 7.2 and 10.1 kb long transcripts. The longest variants of the human and mouse UTRs contain AT-rich sequences (Fig. 3) which have been shown to be involved in mRNA destabilization and degradation (19). The most 3'-end of the longest variant of both genes detected in databases a rat EST derived from a polyadenylated cDNA clone (GenBank accession no. AI044055), which extends the cDNA sequence ~250 bp to another possible poly(A) addition site (Fig. 2b). This site has a GATAAA sequence [~15% polyadenylation activity (17)].

To analyse the expression pattern of the long variant of the mouse *Mecp2* gene, a cDNA clone, JFC617, was used as a probe in RNA *in situ* hybridizations on sections of two stages of mouse embryos and sections of mouse postnatal head. A ubiquitous low level of expression is visible in early organogenesis at 10.5 days post-conceptum (d.p.c.) (Fig. 4a and b). At 12.5 d.p.c. a low level of expression is detectable in the myelencephalon and spinal cord (Fig. 4d and e). At 1 week postnatalis (w.p.n.) expression is visible in all parts of the differentiating brain and in the nasal epithelium, enhanced expression

```
                        3' end of short MeCP2 transcript
            ------------------------------------------------┐AAAAAAAA
                                                            ┌─┘
human       AAGCTTCCCGATTAACTGAAATAAAAAATATTTTTTTTTCTTTCAGTAAA  bp 1669
            |||||||||||||||||||||||||||||||||||||||||||||||||
mouse       AAGCTTCCCGATTAACTGAAATAAAAAATATTTTTTTTTCTTTCAGTAAA  bp 1623
            HindIII          PolyA signal               └─┐
                                                       AAAAAAAAA

human       CTTAGAGTTTCGTGGCTTCAGGGTGGGAGTAGTTGGAGCATTGGGGATGT
            ||||||||||||||||||||| |||||||||||||||||||| |||||||
mouse       CTTAGAGTTTCGTGGCTTCGGGGTGGGAGTAGTTGGAGCATT.GGGATGT

human       TTTTCTTACCGACAAGCACAGTCAGGTTGAAGACCTAACCAGGGCCAGAA
            ||||||||||||||||||||||||||||||||||||||||||||||||||
mouse       TTTTCTTACCGACAAGCACAGTCAGGTTGAAGACCTAACCAGGGCCAGAA

human       GTAGCTTTGCACTTTTCTAAACTAGGCTCCTTCAACAA...    ...
            |||||||||||||||||||||||||||||||||||||||      //
mouse       GTAGCTTTGCACTTTTCTAAACTAGGCTCCTTCAACAA...    ...

human       AGGCAATTTATTAAGGAAATTTGTACCATTTCAGTAAACCTGTCTGAATG  bp 10006
            ||||| ||||||||||||||||||||||||||||||||||||||||||||
mouse       AGGCAATTTATTAAGGAAATTTGTACCATTTCAGTAAACCTGTCTGAATG  bp 9998

human       TACCTGTATACGTTTCAAAAACACCCCCCCCCCACTGAATCCCTGTAACC
            ||||||||||||||||||||||||||| |     ||||||||| |||||||||
mouse       TACCTGTATACGTTTCAAAAACA....CACCCCACTGAACCCCTGTAACC

                        3' end of long MeCP2 transcript
            ------------------------------------------------┐
                                                     AAAAAAAAA
                                                     ┌─┘
human       TATTTATTATATAAAGAGTTTGCCTTATAAATTTACATAAAAATGTCCGT
            ||||||| || ||||||||||||||||||||||||||| |||||||||||
mouse       TATTTATTATATAAAGAGTTTGCCTTATAAATTTACATAAAAATGTCCGT
                                                     └─┐
                                                 AAAAAAAAA

human       TTGTGTCTTTTGTTGTAAAAATCAAGTGATTTTTTCATAAGGTTCTTTTA
            |||||||||||||||| ||||||||| ||||||||||||||||||||||
mouse       TTGTGTCTTTTGTTGT.AAAATCAAGT.GTTTTTCATAAGGTTCTTTTA
```

Figure 3. Sequence comparison of the ends of the 3'-UTRs of the human and murine *MeCP2* genes. The two slashes flanked by periods indicate internal sequences not shown in this comparison. The poly(A) addition sites are indicated. Polyadenylation signals are boxed, GT-rich sequences downstream of the poly(A) addition sites are underlined and three representatives of AT-rich elements ATTTA are double underlined. Nucleotide numbering corresponds to the cDNA sequences (EMBL accession nos AJ132917 and AJ132922).

is found in the olfactory bulb and the hippocampus formation (Fig. 5a and b). In the fully differentiated brain the same expression pattern is found. Enhanced expression is seen in the hippocampus formation and low expression in all other parts of the brain (Fig. 5d and e).

Distribution of conserved blocks and possible three-dimensional structures in the human and mouse MeCP2 transcript

The human and mouse 3'-UTRs of the long variants of the *MeCP2* gene were analysed for the presence of ORFs. The largest human ORFs found in each of the possible forward frames of the human transcript are 459, 273 and 252 bp. None of them is conserved along its entire length between the two species.

To assess which regions of the UTRs might be functionally important, we studied in detail their evolutionary conservation over the whole length. The overall sequence identity between

the human and mouse *MeCP2* long 3'-UTRs is 52%. The homology is interrupted multiple times and extensive gapping is required to maintain the alignment of the two sequences. In contrast to this, the coding region is highly conserved over the whole length. In the 3'-UTR, at least eight blocks of strong sequence similarity are detected. These highly conserved regions are scattered throughout the whole transcript and interrupted by regions of low homology. The conserved blocks are clearly visible when the sequence homology along the two sequences in a sliding 100 bp window is plotted (Fig. 6a). Besides the conservation of the ORF, the strong conservation within the conserved blocks in the 3'-UTR indicates that these sequences are under strong selection pressure and may therefore represent functionally important domains.

The sequence of any RNA molecule may carry information required for its three-dimensional conformation. To detect possible three-dimensional structures, a computer folding of the
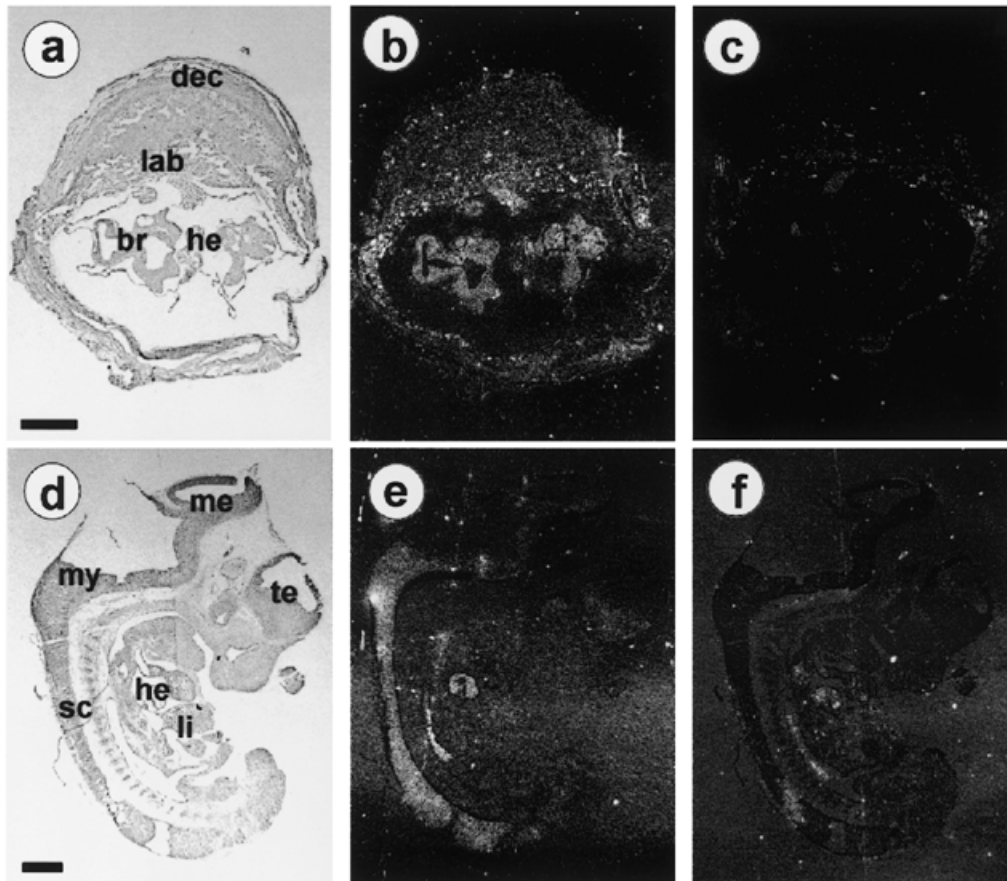
Figure 4. Expression of JFC617 in mouse embryogenesis. Bright field (a and d) and dark field (b, c, e and f) images of a horizontal (a–c) and a sagittal (d–f) section through a 10.5 d.p.c. (**a–c**) and a 12.5 d.p.c. (**d–f**) embryo hybridized with a JFC617 antisense (b and e) and a JFC617 sense (c and f) riboprobe are shown. No specific signals are obtained with the sense control hybridization (c and f). The signals in the heart of the 12.5 d.p.c. embryo are due to non-specific light reflection of embryonic blood cells (e) and are also visible with the sense control probe (f). br, brain; dec, decidua; he, heart; lab, labyrinth; li, liver; me, mesencephalon; my, myelencephalon; sc, spinal cord; te, telencephalon. Bar, 1 mm.

human and murine *MeCP2* sequences was performed using Foldsplit, a program which finds a conformation of minimum free energy using published values of stacking and destabilizing energies (20). Regions of high minimum free energy are expected to have a weak secondary structure. The minimum free energy corresponding to the optimal secondary structure for every 100 bp sequence window was plotted (Fig. 6b and c). Although the overall sequence homology between the human and murine 3' sequences is only 52%, the distribution of the minimum free energy along the sequence is very similar in both. This may indicate that the evolutionary changes did not alter the secondary structure of the RNA molecules. Moreover, the 3'-UTR sequences that show a high degree of conservation during evolution are those regions predicted to be of high free energy with respect to RNA folding (Fig. 6a–c). Interestingly there is no correlation between the degree of conservation and high free energy within the *MeCP2* coding region.

Additional support for a domain-like conservation pattern of the long 3'-UTR of the *MeCP2* gene was obtained from examining conservation in other organisms. To do this, two primers (DC4 and DC5; Fig. 6a) for two neighbouring highly conserved domains were designed. Amplification of chimpanzee, orangutan, macaque, hamster, rat and kangaroo DNA resulted

in the generation of a single PCR product in each species. The size of the products varies between 790 and 846 bp and sequence homology to the human sequence was between 99% (chimpanzee) and 13% (kangaroo). The alignment of all sequences (Fig. 7) was used to generate a phylogenetic tree (Fig. 8). The topology of the tree is in good agreement with the accepted phylogenetic relationships of the species used.

DISCUSSION

Our search for transcribed sequences in Xq28 led to the identification of several long variants of the 3'-UTR of the *MeCP2* gene. The *MeCP2* gene thus belongs to a growing family of genes with multiple tandem terminal poly(A) addition sites, which can be alternatively used in different tissues or developmental stages of an organism (21,22). The phenomenon may be even more frequent and remain partly undetected because alternative transcripts that differ in size by several hundred nucleotides may not be resolved on northern blots because they are contained in the major band, because of the existence of some mRNA variants only in particular cell types or because of their extremely low frequency.
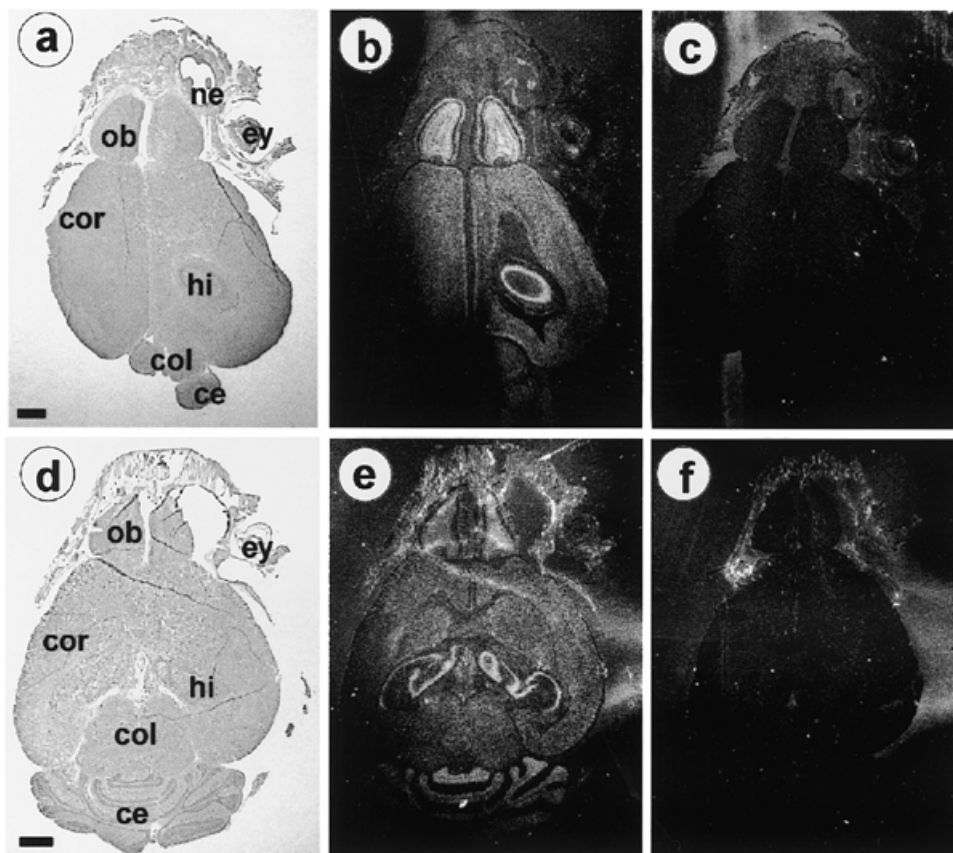
**Figure 5.** Expression of JFC617 in postnatal brain differentiation. Bright field (a and d) and dark field (b, c, e and f) images of horizontal sections through a 1 w.p.n) (a–c) and a 6 w.p.n. (d–f) head hybridized with a JFC617 antisense (b and e) and sense (c and f) riboprobe are shown. Only non-specific signal distribution is visible with the sense control hybridizations (c and f). ce, cerebellum; cor, cortex; col, colliculus; ey, eye; hi, hippocampus; ne, nasal epithelium; ob, olfactory bulb. Bar, 1 mm.

Different distributions of polyadenylation and other factors in particular tissues or developmental stages may direct the abundance of various mRNA species with different 3'-UTRs, which differ in length and content of regulatory elements and may exert different regulatory functions on expression of the gene (12). Some of the variant alternative 3'-UTRs of the *MeCP2* gene show polyadenylation signals that differ from the AATAAA or ATTAAA consensus, similar to other alternatively polyadenylated genes such as *DHFR* and *ATM* or many 3'-EST clusters (22–24). The variants of alternatively polyadenylated genes also differ in the content of AT-rich sequence motifs, which have been implicated in mRNA stability (19). Indeed, there are differences between the *MeCP2* mRNA transcripts in the content of these AT-rich elements. While there is none in the 1.8 kb transcript, many are found close to the 3'-end of the 10.1 kb transcript. The relative abundance of the mRNA isoforms seen on northern blots may also reflect this.

The importance of 3'-UTRs for the regulation of gene expression is increasingly recognized. The 3'-UTRs may exert post-transcriptional regulation enabling rapid modulation of protein levels in response to different developmental or physiological changes and stimuli. This can be especially important during the maturation of germ cells and early embryogenesis, when translationally repressed maternal mRNAs are activated and active mRNAs are silenced (12). During this period there

is no transcription of embryonal genes and all key developmental events are dependent only on proteins and mRNAs which had been formed during oogenesis and which require sophisticated regulation at the mRNA level (25). In the case of the *MeCP2* gene and its product this period may be of particular interest, because it overlaps with the period of gross changes in DNA methylation (26,27) and thus also changes in demand for the *MeCP2* gene product.

Cloning of the complete human and murine transcript of the *MeCP2* genes enabled us to analyse the conservation of these sequences during evolution. The overall conservation of the two 3'-UTRs was 52%, which was lower than the 69.4% average reported for 3'-UTRs between human and mouse (28). Remarkable differences in the degree of conservation could be detected, however, along the transcript. The strong conservation of several sequence blocks within the transcript indicates that these sequences are under selection pressure and therefore functionally important. Some of these sequences have remained without a single base pair change over an evolutionary period of many tens of millions of years since the human and mouse lineages diverged. In contrast to this, other sequences of the *MeCP2* 3'-UTR are only moderately or weakly conserved. This variable pattern of sequence conservation resembles a similar distribution described for several vertebrate 3'-UTRs (29).
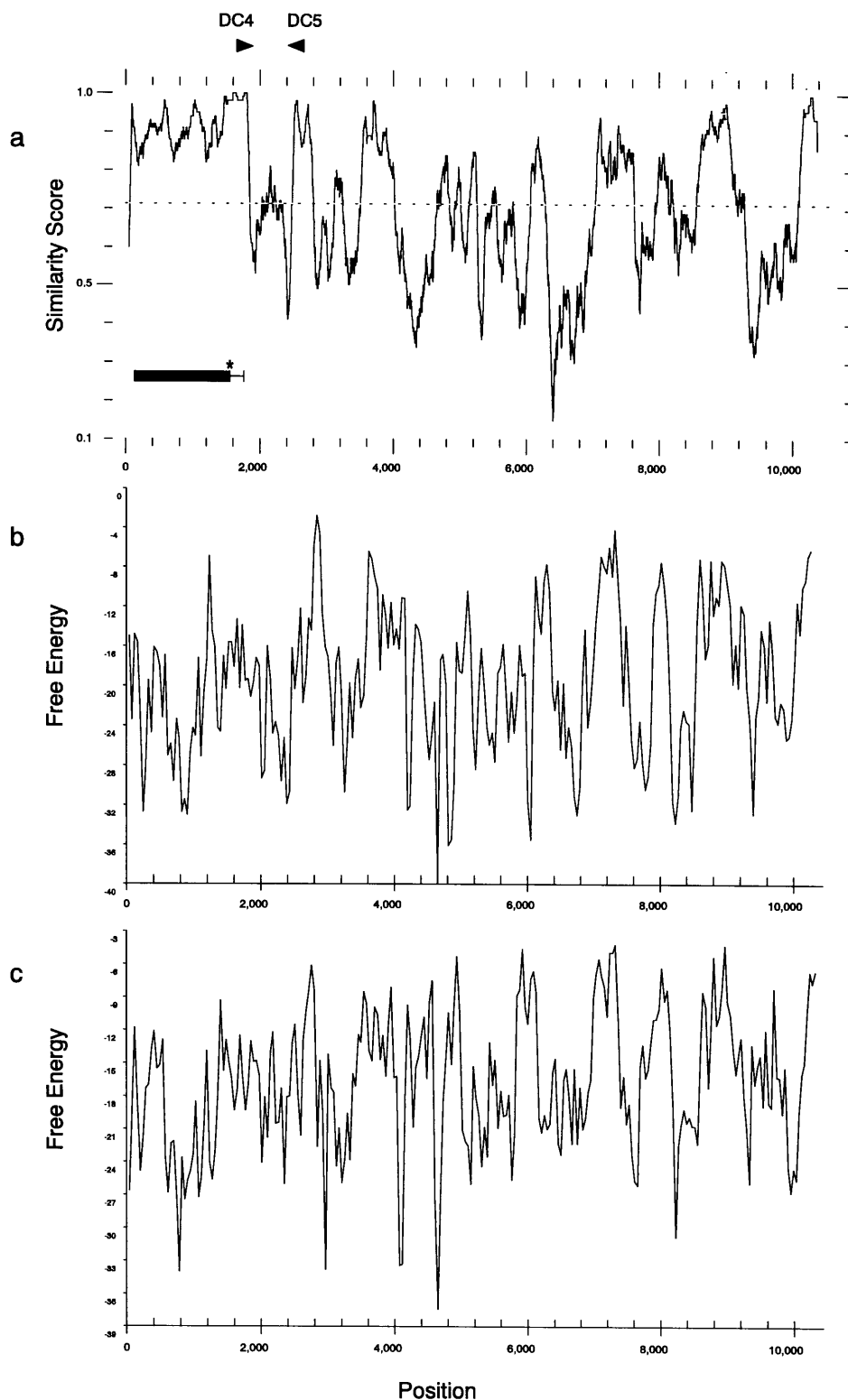
Figure 6. The comparison of the complete sequences of the human and mouse *MECP2* genes. (**a**) A sequence homology comparison was performed using a sliding 100 bp window. (**b**) Minimum free energy plot for the human *MeCP2* mRNA. (**c**) Minimum free energy plot for the mouse *Mecp2* mRNA. The ORF (filled rectangle), stop codon (asterisk), the short 3'-UTR (line) and the positions of primers DC4 and DC5 used for the evolutionary analysis are indicated in (a). The horizontal scale is in bp; the free energy values are according to Zuker and Stiegler (20).

To analyse whether the pattern of conservation is related to predicted three-dimensional structure of the folded RNA molecule, a computer method was used to find a distribution of minimum free energy using published values of stacking and destabilizing energies (20). It was found that the distribution of minimum energy along the 3'-UTR of the transcript is very

```
          1
human    TTGCTGCAGATACTACTGACCAGACAAGCTGTTGACCAGGCACCTCCCCTCCCGCCCAAACCTTT...................CCCCCATGTGGTCGT
chimp    ----------------------------------------------------------------....................---------------
orang    ----------------------------------------------------------------....................---------------
macaq    ----------------------------------------------------------------....................---------------
hamst    ---------------------------------------------.,.---C..-AATA...-TC...................---A-----C--A-
mouse    ------------------------------------------TC----CAA-AATATCCT-CC-CTTCCCCCCCCCCACCCCCG----.G----C----
rat      ------------------------------------------TC---ACAA-AA-A-CC--C-CCCTCCTCACCCCACCCCTAT----TG----C--A-
kanga    --------T--------------------T-T-----TTT-TAGGGTA-A--AGC----GC....................--T----TCATC-C--

          101
human    TAGAGACAGAGCGACAGAGCAGTTGAGAGGACACTCCCGTTTTCGGTGCCATCAGTGCCCCGTCTACA...GCTCCCCCAGCTCCCCCC..ACCTCCCCC
chimp    ------------------------------------------------------------------...---------------...----------
orang    ------------------.......------.-----------------------------------...---G--.-----------...----------
macaq    ------------------------------------------------------------------C..-A------------...----------
hamst    ------..T---.......-------------------A---T------.......--..T---C-T-...--T---TGA---TT--...--A----A
mouse    ---..-----.G--....-------A----------A---T------TGA------T---C-T-ATA---T---TGA--TTTA-...--A----A
rat      ------.--G---.......-A----------A---T------CTGA-----T---C-T-...---T---TGA--TTTA-...--A----A
kanga    ..................----G---AGGGA-GGT--AC-G-T--T--CT--A-A--TAA-AGAGTAG-G-TAGTGG-AG--TTA-ATTTT-AGTG--

          201
human    ACTCCCAACCACGTT.GGGACAGGGAGGTGTGAGGCAGGAGAGACAGTT..GGATTCTTTAGAGAAGA...TGGATATGACCAGTGGCTATGGCCTGTGC
chimp    ---------------.------------------------------.------------------...---------------------T
orang    ---------------.-------------------------------.---------C-------...---------------C-----------
macaq    ---------------.------------------------------.---------------------G--------------------C----
hamst    --------T-TGA...------C---------C---------A-A-T-T-G..----------------CTA...--GG--------A--G----CA.-T
mouse    --------T-TGA...A----T---------.-T--------A--TA-G..--C----GG-------CTA----GT--G-----AT--A----CA.-T
rat      --------T-TGA...-----T----------C--------A--TA-ATA---C----GG-------CTA-A--GT--G-A----AT-GC-C--CA.-T
kanga    -T-AA--TT-TA--CCAA-GTCTTA-AT-A-T-T-TT-AG-G-TTTT--........---CCCT---GG-GCC.G--GGG--GGGG-A--G-.--ATTA-...

          301
human    GATCCCACCCGTGGTGGCTCAAGTCTGGCCCCACACCAGCCCCAATCCAAAACTGGCAAGGACGCTTCACAGGACAGGAAAGTGGCACCTGTCTGCTCCA
chimp    ----------------------------------------------------------------------------------------------
orang    --------------C-------------------------------------------------------------------------------
macaq    --------------C-------A----------C--------------------------------.-------------------------T---.
hamst    A--TAG-A-T--------A--------------T-CA---.------------AT----TAT--TGA-AA---------CA-T---------A----
mouse    A--T---A-T-----A--A--GA------T----T--A---.----------A------TAT--TG--AA-A-A-----------------A----
rat      A--T---A-T-----A--A---------T-------A---.-----------A------AT--TG--AA-A-T----------TT------A----
kanga    AT---TTAGGAAA-A--G-TG--A-A-A-AG--G-G--CTGAGC-GTTGGC---A.-----GA-C-TGACTAGGG-CC-G------T-.A---AA---.

          401
human    GCTCTGGCATGGCTAGGAGGGGGGAGTCCCTTGAACTACTGGGT.GTAGACTGGCCTGAACCACAGGAGAGGATGGCCCAGGGTGAGGTGGCATGGTCCA
chimp    ------------------------------------------.----------------------------------------------G-----A
orang    --------------------A-TC------------------------------------------A--T------------------------
macaq    ....----------------A-TT-------------------.------------A--T------------------------------
hamst    -------T--A-G-------T-A-T....---------G----A-TAC-------T---G------A---T-----A---A-A-A--AT---CAC-TG
mouse    ------A----------AG-T-A-T....-.-A----GA--C-TA--A---A-------G----A---T--------A---A---T---CA--TG
rat      ---------------AGAT-A-T....-T.-A----GT---C-TA--A----------G-A----A-----------A--A-A---T---CA--TG
kanga    .....CAAGG----CCA--T-A-T.....A--AGGG-GGG--AAGAC-T-A-A--AAGGA-TAG..AA-C-----A--T---CC-A-A-AA-AGC---T..

          501
human    TTCTCAAGGGACG.TCCTCCAACGGGTGGCGCTAGA.....GGCCATGGAGGCAGTAGGACAAGGTGCAGGCAGGCTGGCCTGGGGTCAGGCCGGGCAGA
chimp    --------------.------------------....-------------------------C---------------C----------------
orang    --------------.--------------AA...-----------------------C-----------------C------------------G
macaq    --------------.------GTT------A-----GA...-----------------CA-----------CA----------------C--
hamst    C--A------C-T-C-T-A-T-GA-AA-AATT---A-GAGG.T-----------C---------CA-A----T---GTG------A-A--CT-
mouse    ---A------C-T-C---C-T-GAA-A-AAT----A-GAGG.T-----------C-------A--A---------A-GTG-A-.---A---A---CT-
rat      ---A------C-T-C---C-T-GAA-T-CAT----A-GAAG.T-----------C------A--A-----.....-A-GTG-A-.---A---A---CT-
kanga    .....--AACC-TAC--A---GGA-T--A-TTG-A-GAGGCCC-T-------A-TCCCC-ACCACCAA-A---....AT-T--A-C-GCA--T..--TTC

          601
human    GCACAGCGGGGTGAGAGGGATTCCTAATCACTCAGAGCAGTCTGTGACT.............TAGTGGACAGGGGAGGGGGCAAAGGGGGAGGAGAAGA
chimp    -----------------------------------------------------------------------------------------------
orang    ---------------------------------G------GGTAGTTAGGGACTC----------------.-G-G----C---------
macaq    A---------------------CGT-.------------------CGGTAGTTAGGGACT-----------A-----.-----------------
hamst    -TG-CA-A-AACA----G.TG---...-T---TA-G---...........--T-----TG--AC-A---.GC-T------A--A-T--
mouse    -TG-CA-A-AACA----A-CAGT--...G---...--T-A-TAAGA...............................A--A-G--
rat      -TG-CA-A-A-CA----A-CAGT--C.TG---...--T---TAAG-.................................----A--A-GA-
kanga    ATTT--T--ACCTT-T-TATA--TGGGTGTG-ATGCAC-TAGATA---.................................A-TGA--A-GA-

          701
human    AAATGTTCTTCCAGTTACTTTCCAATTCT...CCTTTAGGGACAGCTTAGAATTATTTGCACTATTGAGTCTTCAT....GTTCCCACTTCAAAACAAAC
chimp    ---------------------------...------------------------------------------....-----------------
orang    ---------------------------...-----------------------------------------....-----------------
macaq    ------------------------ACT-------------------------------------....----------------
hamst    -----A--------A----------G----...--------------------------------------A---------
mouse    ----A--------A----------G----...--------------------------------------------------
rat      ----A--------A--G-------G----...--------------------------------------------------T
kanga    --C------------CT------G-G--ACTAGC---------G------C-G-C------A----T--GA----TCCCA------------------

          801
human    AGATGC.....TCTGAGAGCAAACTGGCTTGAATTGGTGACATTTAGTCCCTCAAGCCACCAGATG......TGACAGTGTTGAGAACTACCTGGATTTG
chimp    ------.....-------------------------------------------------.....------------------------------
orang    .------.....-------------------------------------------------.....------------------------------
macaq    -----.....------------------------G----------G-----------.....---TG----------------A--
hamst    ------....-----A--------------A--------C--T------A---------A---......--G------T----------C--
mouse    ------....-----A--------------A--------C--.-----A----------CA......-G-------C---------T--C--
rat      ------....-----A--------------A--------C-..-----A----------CAA.....-G-------C-----------T--A--
kanga    T--CTGAGATG-TCA-C--A---------C---G---A----GCCC-TG-CACTTA-TTA---CACTGGTG---TG-G---TT-----C---AT-----

          901
human    TATATATACCTG
chimp    ------------
orang    ------------
macaq    ------------
hamst    ------CCTG..
mouse    ------......
rat      ------CCTG..
kanga    --GG---CTG..
```

Figure 7. Sequence comparison of fragments of human, chimpanzee, orangutan, macaque, hamster, mouse, rat and kangaroo *MeCP2* gene 3'-UTRs. Identities are indicated with a dash, gaps with a dot; different bases are shown. Numbering is relative, with human base 1 here corresponding to base 1811 of the human sequence and murine base 1 corresponding to base 1764 of the mouse sequence.
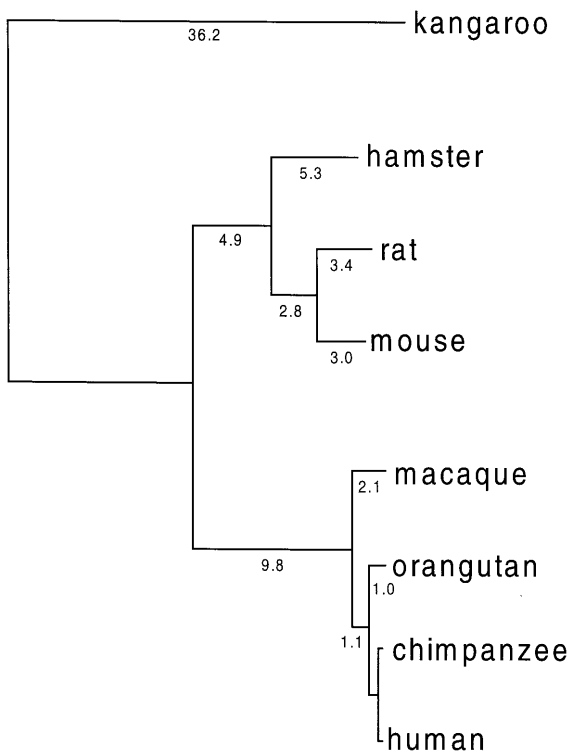
Figure 8. Phylogenetic tree based on the alignment shown in Figure 7.

similar in human and mouse. This may point to a conserved three-dimensional structure of the 3'-UTR molecule. Moreover, there is a correlation between distribution of minimum free energy and the degree of sequence conservation in the 3'-UTR. The sequences that show the highest degree of conservation show the highest level of minimum energy and are likely to be unstable with respect to three-dimensional folding and therefore single stranded. The sequences that are less conserved during evolution are those that show low levels of minimum free energy and therefore are likely to build well-packed stable three-dimensional structures.

We conclude that the sequence of the *MeCP2* 3'-UTR can be divided into two subtypes. First, conserved regions where the primary structure (nucleotide sequence) itself is important. Second, sequences that are less conserved but may build stable and conserved three-dimensional structures. Both types of conservation, the conservation of sequence and of three-dimensional structure, could be of relevance for possible interactions with protein or other nucleic acid molecules. These may exert regulatory functions and modulate the stability, transport, localization and translation of the mRNA (12). The isolation and characterization of the complete long 3'-UTRs of the human and mouse *MeCP2* genes opens the possibility of studying these phenomena on the functional level.

Because of the periodic occurrence of conserved blocks (which allow placement of primers) and stretches of weakly conserved sequences (from which valuable phylogenetic information can be obtained), the 3'-UTR of the *MeCP2* gene may also be very useful in evolutionary studies. We isolated one such highly diverged region from chimpanzee, orangutan, macaque, rat, hamster and kangaroo and compared it with the human and mouse sequences. Indeed, the phylogenetic rela-

tions of the sequences reflect the currently accepted phylogenetic relations between these organisms.

MATERIALS AND METHODS

Isolation of cDNA and genomic clones, sequencing and sequence analysis

Plaques (2×10^6) from each of the human and mouse fetal brain cDNA libraries (Stratagene, La Jolla, CA) were screened with the different cDNA and genomic probes. DNA probes were labelled with [α-^{32}P]dATP and [α-^{32}P]dCTP (3000 Ci/mmol) in a random primed reaction (30). Hybridizations were carried out in 0.5 M sodium phosphate, 7% SDS, 0.2% bovine serum albumin, 0.2% PEG 6000, 0.05% polyvinylpyrrolidone 360000, 0.05% Ficoll 70000 and 0.5% dextran sulphate at 65 or 55°C overnight. Non-specifically bound probe was removed by washing at 65 or 55°C in 40 mM sodium phosphate, pH 7.2, 1% SDS for 60 min. Filters were exposed to X-ray film (XAR-5; Kodak, New York, NY) for 1–5 days. Multiple human fetal and adult tissue poly(A)$^+$ RNA northern blots were purchased from Clontech (Palo Alto, CA). A total mouse P1 library (16) was screened with the cDNA probes JFC88 and JFC484.

For sequencing, the dideoxy termination method was used (31). Incorporation of the dideoxynucleotides into the sequencing product was improved by the use of manganese-containing buffer (32).

Multiple and pairwise alignments of DNA and protein sequences were performed using Clustal V (33) and Bestfit (34), respectively. The percentages of identity and similarity of sequences were calculated using Bestfit. Database searches employed FastA (35). All sequence analysis used the HUSAR 3.0 Sequence Analysis Package (DKFZ, Heidelberg, Germany). The sequences reported in this paper have been deposited in the EMBL databank with sequence accession nos AJ132914–AJ132924.

PCR and RT–PCR

Total RNA of EBV-transformed human lymphocytes and RNA of different cell lines (SK-N-SH, KELLY and IMR-32; ATCC, Manassas) was isolated (36). First strand cDNA was synthesized using 250 ng hexanucleotides/μg total RNA. In addition, mouse cDNA libraries (Stratagene and Clontech) and total genomic DNA from different species were used as templates for PCR amplifications.

The PCR reactions with first strand cDNA, cDNA library aliquots and total genomic DNA comprised 35 cycles of 1 min at 94°C, 1 min at 60°C and 4 min at 72°C.

The following primers were used: DC4, aaactaggctcct-tcaacaaggc, and DC5, tgagcccactttaaaacaagcg (conserved primers in the 3'-UTR for amplification in different species); MEC1, tgttagggctcagggaagaaaag, and MEC3, gggaagctttgtca-gagccctac (for amplification of the human coding region); MECP-CON-1, tgacatgtgactccccagaatac (located in the 5'-UTR and conserved between human and rat), and MECP-MOU-4, ctgctgctcaagtcctgaggctc (for amplification of the mouse coding region; designed from a mouse EST, GenBank accession no. AA690741); MEC2, gtgtatttgatcaatccccaggg, and D44, tccttgccagtttggattgggg (for connection of the short *MeCP2* transcript with the contig of novel cDNAs in human);

MEC-MOU-11, aggaccccatcagcccccctgag, and MEC-MOU-18, cagcttgtcaggtcagtagtatc (as above for mouse).

The isolated cDNA clones were as follows. Human: JFC2, bp 1769–3487 (agtagctttgca...tgctgtgggttt); JFC88, bp 3258–6627 (cagcagggc...cgacttgtg); JFC208, bp 3758–7159 (gtccct-ggc...taaaaaaag); JFC487, bp 6920–9304 (cctgggctt...aacact-gtg); JFC484, bp 7237–10084 (cttctttca...ttgccttat). Mouse: JFC611, bp 3022–7140 (tctgggtca...gaaaaaaaa); JFC617, bp 3452–7152 (tcatattgg...aagaaaaaa); JFC622, bp 7762–10087 (tatctgcta...acataaaaa).

RNA *in situ* hybridization

The mouse embryo and adult tissue sections were prepared and the RNA *in situ* hybridizations were performed as described previously (37). The probe used in these experiments was the 3.7 kb JFC617 mouse cDNA insert transcribed *in vitro* from the pBluescript II vector promoters.

ACKNOWLEDGEMENTS

REFERENCES

1. Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurer-Fogy, I., Jeppesen, P., Klein, F. and Bird, A. (1992) Purification, sequence and cellular localisation of a novel chromosomal protein that binds to methylated DNA. *Cell*, **69**, 905–914.
2. Bird, A.P. (1986) CpG rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
3. Siegfried, Z. and Cedar, H. (1997) DNA methylation: a molecular lock. *Curr. Biol.*, **7**, R305–R307.
4. Nan, X., Ng, H.-H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N. and Bird, A. (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, **393**, 386–389.
5. Jones, P.L., Veenstra, G.J.C., Wade, P.A., Vermaak, D., Kass, S.U., Landsberger, N., Strouboulis, J. and Wolffe, A.P. (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genet.*, **19**, 187–191.
6. Tate, P., Skarnes, W. and Bird, A. (1996) The methyl-CpG-binding protein MeCP2 is essential for embryonic development in the mouse. *Nature Genet.*, **12**, 205–208.
7. Li, E., Bestor, T.H. and Jaenisch, R. (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.
8. Coy, J.F, Kioschis, P., Sedlacek, Z. and Poustka, A. (1994) Identification of tissue-specific expressed sequences in human band Xq28 with complex pig cDNA probes. *Mamm. Genome*, **5**, 131–137.
9. Coy, J.F., Dübel, S., Kioschis, P., Thomas, K., Micklem, G., Delius, H. and Poustka, A. (1996) Molecular cloning of tissue-specific transcripts of a transketolase related gene: implications for the evolution of new vertebrate genes. *Genomics*, **32**, 309–316.
10. Coy, J.F., Velhagen, I., Himmele, R., Delius, H., Poustka, A. and Zentgraf, H. (1996) Isolation, differential splicing and protein expression of a DNase on the human X chromosome. *Cell Death Differ.*, **3**, 199–206.
11. D'Esposito, M., Quaderi, N.A., Ciccodicola, A., Bruni, P., Esposito, T., D'Urso, M. and Brown, S.D (1996) Isolation, physical mapping and northern analysis of the X-linked human gene encoding methyl CpG-binding protein, MeCP2. *Mamm. Genome*, **7**, 533–535.
12. Decker, C.J. and Parker, R. (1995) Diversity of cytoplasmic functions for the 3' untranslated region of eukaryotic transcripts. *Curr. Opin. Cell Biol.*, **7**, 386–392.
13. Kioschis, P., Gong, W., Rogner, U.C., Wilke, K., Manca, A., Coy, J.F. and Poustka, A. (1994) Cosmid contigs in Xq27.3–Xqter. *Cytogenet. Cell Genet.*, **67**, 355.
14. Korn, B., Sedlacek, Z., Kioschis, P., Konecki, D.S. and Poustka, A. (1994) Establishment of an Xq27.3–Xqter specific cDNA library using cDNA selection. *Cytogenet. Cell Genet.*, **67**, 356.
15. Cao, Z., Henzel, W.J. and Gao, X. (1996) IRAK: a kinase associated with the interleukin-1 receptor. *Science*, **271**, 1128–1131.
16. Zehetner, G. and Lehrach, H. (1994) The reference library system—sharing biological material and experimental data. *Nature*, **367**, 489–491.
17. Wickens, M. (1990) How the messenger got its tail: addition of poly (A) in the nucleus. *Trends Biochem. Sci.*, **15**, 277–281.
18. Proudfoot, N. (1991) Poly(A) signals. *Cell*, **64**, 671–674.
19. Chen, C.-Y.A. and Shyu, A.-B. (1995) AU-rich elements: characterisation and importance in mRNA degradation. *Trends Biochem. Sci.*, **20**, 465–470.
20. Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
21. Edwalds-Gilbert, G., Veraldi, K.L. and Milcarek, C. (1997) Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res.*, **25**, 2547–2561.
22. Gautheret, D., Poirot, O., Lopez, F., Audic, S. and Claverie, J.-M. (1998) Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.*, **8**, 524–530.
23. Hook, A.G. and Kellems, R.E. (1988) Localisation and sequence analysis of poly(A) sites generating multiple dihydrofolate reductase mRNAs. *J. Biol. Chem.*, **263**, 2337–2343.
24. Savitsky, K., Platzer, M., Uziel, T., Gilad, S., Sartiel, A., Rosenthal, A., Elroy-Stein, O., Shiloh, Y. and Rotman, G. (1997) Ataxia-telangiectasia: structural diversity of untranslated sequences suggest complex post-transcriptional regulation of ATM gene expression. *Nucleic Acids Res.*, **25**, 1678–1684.
25. Wickens, M. (1990) In the beginning is the end: regulation of poly(A) addition and removal during early development. *Trends Biochem. Sci.*, **15**, 320–324.
26. Jaenisch, R. (1997) DNA methylation and imprinting: why bother? *Trends Genet.*, **13**, 323–329.
27. Kass, S.U., Pruss, D. and Wolffe, A.P. (1997) How does DNA methylation repress transcription? *Trends Genet.*, **13**, 444–449.
28. Makalowski, W. and Boguski, M.S. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA*, **95**, 9407–9412.
29. Duret, L., Dorkeld, F. and Gautier, C. (1993) Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Res.*, **21**, 2315–2322.
30. Feinberg, A.P. and Vogelstein, B. (1983) A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.*, **132**, 6.
31. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, **74**, 5463–5467.
32. Tabor, S. and Richardson, C.C. (1987) DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl Acad. Sci. USA*, **84**, 4767.
33. Higgins, D.G. and Sharp, P.M. (1988) CLUSTAL, a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.
34. Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, **12**, 387–395.
35. Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
36. Chomczynski, P. and Sacchi, N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. *Anal. Biochem.*, **162**, 156–159.
37. Bächner, D., Sedlacek, Z., Korn, B., Hameister, H. and Poustka, A. (1995) Expression patterns of two human genes coding for different rat GDP-dissociation inhibitors (GDIs), extremely conserved proteins involved in cellular transport. *Hum. Mol. Genet.*, **4**, 701–708.