Bioinformatics and Biology Insights

ORIGINAL RESEARCH

Comparative *in Silico* Analyses of *Cpeb1–4* with Functional Predictions

Xiang-Ping Wang and Nigel G.F. Cooper

Anatomical Sciences and Neurobiology, University of Louisville, Louisville, KY 40292, USA.
Corresponding author email: nigelcooper@louisville.edu

Abstract

Background: Cytoplasmic polyadenylation element binding proteins (Cpebs) are a family of proteins that bind to defined groups of mRNAs and regulate their translation. While Cpebs were originally identified as important features of oocyte maturation, recent interest is due to their prospective roles in neural system plasticity.

Results: In this study we made use of bioinformatic tools and methods including NCBI Blast, UCSC Blat, and Invitrogen Vector NTI to comprehensively analyze all known isoforms of four mouse *Cpeb* paralogs extracted from the national UniGene, UniProt, and NCBI protein databases. We identified multiple alternative splicing variants for each *Cpeb*. Regions of commonality and distinctiveness were evident when comparing *Cpeb2, 3*, and *4*. In addition, we performed cross-ortholog comparisons among multiple species. The exon patterns were generally conserved across vertebrates. Mouse and human isoforms were compared in greater detail as they are the most represented in the current databases. The homologous and distinct regions are strictly conserved in mouse Cpeb and human CPEB proteins. Novel variants were proposed based on cross-ortholog comparisons and validated using biological methods. The functions of the alternatively spliced regions were predicted using the Eukaryotic Linear Motif resource.

Conclusions: Together, the large number of transcripts and proteins indicate the presence of a hitherto unappreciated complexity in the regulation and functions of Cpebs. The evolutionary retention of variable regions as described here is most likely an indication of their functional significance.

Keywords: *in silico*, *Cpeb*, bioinformatics, isoforms, paralogs, orthologs, alternative splicing

Bioinformatics and Biology Insights 2010:4 61–83

This article is available from http://www.la-press.com.

Introduction

Cytoplasmic polyadenylation element binding proteins are a family of mRNA binding proteins that play essential regulatory roles in the translation of defined mRNAs. First discovered during oocyte maturation,[1] the role of Cpeb-mediated control of translation has now been expanded to include a wider variety of scenarios including cell cycling[2,3] and synaptic plasticity.[4] The identification of *Cpebs* in a wide variety of tissues[5,6] indicates that they may function as a ubiquitous means for controlling the translation of specifically targeted mRNAs.

Four *Cpeb* paralogs have been identified in mouse. The first family member, *Cpeb1,* was identified using single-step RNA affinity chromatography. Enriched in oocyte, it is indispensible for cytoplasmic poly(A) elongation during oocyte maturation.[1] Transcripts for *Cpeb2* were first identified in mouse testis using an EST database and degenerative PCR.[1,7] *Cpeb3* and *Cpeb4* were first detected in mouse brain via PCR and Northern blotting using primers/probes similar to human *CPEB*-like sequences.[5] The N termini of *Cpeb1–4* are highly variable, whereas the C-termini, where RNA recognition motifs (RRMs) reside, are more conservative. Sequence analysis has revealed that *Cpeb1* is distant from *Cpeb2, 3* and *4* in the family tree.[5] Expression of *Cpeb1, 2, 3* and *4* mRNAs in the hippocampus demonstrated overlapping, yet distinct patterns.[8] *Cpeb3*, in particular, has been associated with human memory.[9] The cytoplasmic polyadenylation element (CPE), a short U-rich motif, has been identified in the 3'UTRs of mRNAs targeted by Cpeb1,[10,11] while a distinct loop-forming U-rich motif appears to be indispensible for the binding of Cpeb4 and Cpeb3, but not of Cpeb1 protein.[8]

Previous biological findings suggested that *Cpeb* paralogs, although distinct in their own ways, may share some commonality in their structure and distribution, and may possibly provide some compensation and redundancy in their function. A systematic analysis of *Cpebs* based on the current databases and literature would surely be informative and instructive to ongoing *Cpeb*-related research. The purpose of the current study is to perform a comprehensive survey and analyses on three scales: within each paralog, across-paralog, and across-ortholog. Through data mining of the current nucleotide and protein databases and previous publications, we derived the alternative splicing patterns for each *Cpeb*. Some of the newly proposed alternatively spliced regions were confirmed experimentally. Cross-paralog and cross-ortholog comparisons illuminated the similarities and the unique attributes of four *Cpeb*s, as well as the extraordinarily high level of conservation of each *Cpeb* across species. A bioinformatics analysis revealed the presence of specific functional motifs.

Results and Discussion

Cpeb1 protein isoforms with internal deletions of 1 or 5-amino acid (aa), or with an N-terminal truncation of 75-aa

A total of nine cDNA sequences for mouse *Cpeb1* were extracted from the UniGene database (supplementary Table 1). Fragmented sequences and redundant sequences were identified with the bioinformatics tools Blast and Vector NTI and removed from further analysis. Four non-redundant full-length cDNAs were aligned to mouse genomic DNA (derived from the UCSC mouse genome) to infer exon-exon boundaries and to derive alternatively spliced exons (Fig. 1A). The comparison demonstrated that the variances in the lengths of the first and last exons lead to different 5' UTRs or 3' UTRs, respectively (Fig. 1A). Two variable sequences in the protein coding region (CDS), including a 3-nucleotide (nt) deletion resulting from partial exon 4 skipping and a 15-nt deletion resulting from partial exon 7 skipping, would lead to altered proteins. The presence of transcripts with or without the 15-nt variable region has been confirmed in mouse brain, ovary,[12] and retina (Fig. 1B left).

Two Cpeb1 protein sequences were extracted from the UniProt protein database and aligned using Vector NTI software (Fig. 1C). Meanwhile, we computationally translated all non-redundant full-length *Cpeb1* transcripts with the aid of Vector NTI, and then compared the translated protein products to the protein sequences in the database. Our computational translation of cDNA BC144948.1 yielded a protein with a 5-aa deletion, which is not documented in the protein database (Fig. 1C). The removal of the 5-aa motif is due to partial skipping of exon 7 (15-nt) as previously described (Fig. 1A). In addition to the evidence at the transcript level in the mouse (Fig. 1B left), two isoforms of human CPEB1 proteins with the same 5-aa deletion[13] were

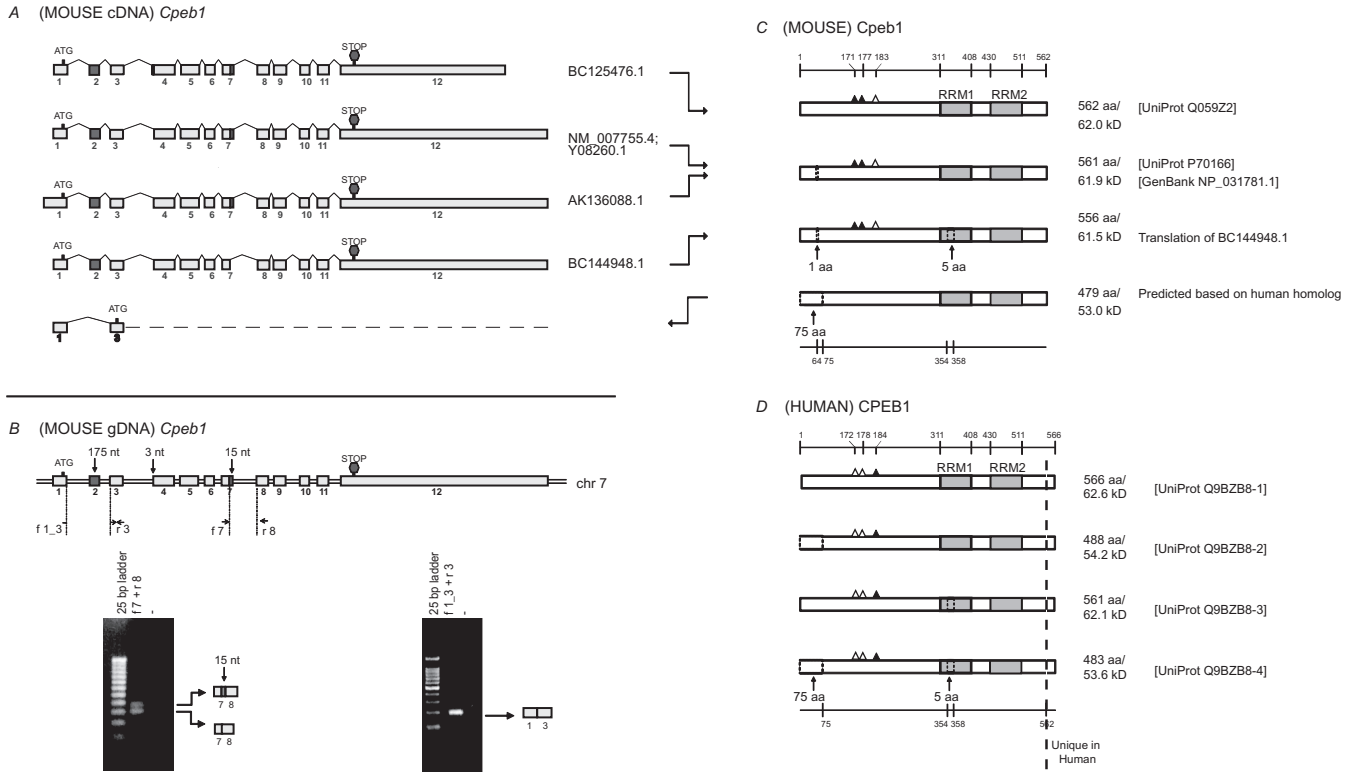


Figure 1. Analysis of *Cpeb1*. **A)** Transcripts of mouse *Cpeb1*. Only non-redundant full-length UniGene sequences were used for this analysis, with their accession numbers listed on the right. ATG and STOP indicate the presence of translational initiation and termination sites, respectively. The different lengths of the first and last exons likely represent the presence of variable 5' and 3' UTRs, respectively. The alternative splices of the first 3-nt of exon 4 and the last 15-nt of exon 7 (highlighted in grey) would generate different protein products. A novel isoform with deletion of exon 2 (also highlighted in grey) is predicted based on published sequences of the human protein. The deletion of exon 2 leads to the use of an alternative translational start codon in exon 3. **B)** Expression of *Cpeb1* transcripts in adult mouse retina. The locations of the primers for RT-PCR are aligned to the diagram of *Cpeb1* genomic DNA, in which boxes represent exons and double lines represent introns. Photographs of DNA gels demonstrate the expression of multiple Cpeb1 transcripts in the retina including those with and without the 15-nt of exon 7 (left), and without exon 2 (right). The identity of each band was confirmed by nucleotide sequencing. **C)** Isoforms of mouse Cpeb1 proteins. Two isoforms were extracted from the UniProt database (Q059Z2, P70166). The computational translation of cDNA BC144948.1 yields a third isoform. A fourth isoform is predicted based on two human CPEB1 homologs (Q9BZB8-2, Q9BZB8-4). RNA recognition motifs (RRMs) are indicated with grey boxes. Triangles represent phosphorylation sites experimentally confirmed (solid)[21] or predicted (open) based on cross-ortholog comparisons. A 1-aa deletion, a 5-aa deletion, and a 75-aa N-terminal truncation are each indicated with dashed line boxes. The locations of functional motifs are shown as numbered amino acid sites at the top of the diagram, and those of the alternative spliced regions at the bottom, as might be seen in the longest isoform. **D)** Isoforms of human CPEB1 proteins. Four isoforms are extracted from the UniProt database. The RRMs, the phosphorylation sites, and the 5-aa deletion are all present in human CPEB1 at the same locations as seen in mouse. Two isoforms have 75-aa N terminal truncations (Q9BZB8-2, Q9BZB8-4). This led to our predication of the existence of a similar isoform in mouse. Human CPEB1 has an additional 4-aa at the C-terminus than mouse Cpeb1 (shown to the right of the dashed line). Numeric annotations refer to the longest isoform.

identified in the UniProt database (Fig. 1D). Stringent homology is evident between mouse and human as we later conclude. In particular, the locations and sequences of the 5-aa deletion in mouse Cpeb1 and human CPEB1 are identical (Fig. 1C, 1D, Fig. 6D). Additional evidence includes the presence of the 90-nt and 105-nt variants of a particular exon, which corresponds to exon 7 in mouse (supplementary Table 2, the asterisk), across vertebrates. The stringent conservation of exon patterns across vertebrates strongly supports the presence of the 15-nt (5-aa) alternative splices within this exon which likely will have important functional implications.

The 15-nt (5-aa) is located within the first RRM (Fig. 1C). Further analysis identified that the 5-aa is adjacent to the octamer consensus of the first RRM (Fig. 5, green box). The insertion or deletion of the 5-aa may have a pivotal impact on the specificity of Cpeb1, because the sequences surrounding the consensus of RRMs are important for the specificity of RNA binding.[14,15] To what extent this 5-aa impacts RNA binding is yet to be established.

Two alternative splicing isoforms containing a 75-aa N-terminal truncation were evident in human CPEB1 (Fig. 1D). With the knowledge that there is a near-identical conservation between human

CPEB1 and mouse Cpeb1 (Fig. 6B–D), the question arises: Does the 75-aa truncation in human have a counterpart in mouse? Based on our theoretical translation with Vector NTI, we postulated that the 75-aa truncation in mouse could be derived from the removal of exon 2, which leads to a frame shift and an alternative translational initiation site (Fig. 1A). Therefore, we designed primers spanning this alternative region in mouse. A primer pair at exon 1/3 junction and within exon 3 confirmed the presence of a *Cpeb1* transcript "exon 2 deletion" in mouse retina (Fig. 1B right). This PCR-based evidence at the level of transcripts provided support for the presence of this novel isoform of mouse Cpeb1 with 75-aa N-terminal truncation.

Cpeb2 protein isoforms with internal deletions of 30-aa or 8-aa

We extracted eight cDNA sequences for *Cpeb2* from the UniGene database (supplementary Table 3). Three non-redundant, full-length sequences were used for further analysis (Fig. 2A). A recently updated sequence for one isoform (NM_175937.3) was also included in the diagram. The alignment demonstrated that the variances in the lengths of the first and last exons of *Cpeb2* lead to different 5' UTRs and 3' UTRs, respectively. The partial skipping of exon 2 removes 3-nt from the 5' UTR. The alternative splices of exon 4 and exon 7 remove 90-nt and 24-nt respectively from the coding region. Our RT-PCR results confirmed the expression of transcripts with or without the 90-nt variable region in adult mouse retina (Fig. 2B). Transcripts without the 24-nt was not detected, perhaps due to competition for the same primers which can lead to masking of the less abundant isoforms by the more dominant ones, as observed in *Cpeb3*.[6] Alternatively, this could be due to a distinct tissue specificity and/or condition. The alternative use of these two regions was also observed in a number of other vertebrate species (supplementary Table 4, the asterisks).

One mouse Cpeb2 protein sequence was documented in the UniProt database (Q812E0). An additional isoform which was recently updated in the NCBI protein database (but not in the UniProt database) was added to the top (Fig. 2C, NP_787951.2). Both sequences contain an 8-aa deletion resulting from the removal of exon 7 (Fig. 2A). When we translated non-redundant full-length *Cpeb2* transcripts using Vector NTI, we identified a novel protein from the translation of cDNA AK0421065.1 (Fig. 2C). This predicted isoform has a 30-aa deletion resulting from the removal of exon 4 (90-nt), which has been confirmed at the level of transcripts (Fig. 2B). In addition, we identified the same 30-aa deletion in human CPEB2 (Fig. 2D). Six human CPEB2 protein isoforms, including the one recently deposited in the NCBI protein database (NP_001170853.1), are distinguished by the presence or absence of three motifs of 22-aa, 30-aa, or 8-aa each (Fig. 2D). The sequences and the relative locations of the 30-aa and 8-aa in human CPEB2 are comparable to those in mouse. The strong homology between mouse and human (Fig. 6B–D) and the similar findings in human provided additional support for the presence of a mouse Cpeb2 isoform with a 30-aa deletion.

Both mouse and human Cpeb2 sequences have been updated in the NCBI database during the preparation of this manuscript. Of particular interest, the updated sequences demonstrate the presence of an extra-long isoform of Cpeb2—almost double the previously published size (Fig. 2C, 2D, top isoforms). Our sequence alignment demonstrated that both the previous and the updated mouse isoforms are legitimate—the use of an extended exon1 leads to the much longer N terminus in the newly uncovered isoform (Fig. 2A, insert). This may be of relevance to prior investigations of CPEB3, in which antibodies recognized the predicted protein at ~78 kD in western blots, but also detected an additional protein band above 100 kD[6,8] (see below).

Cpeb3 protein isoforms with internal deletions of 23-aa or 8-aa, an N-terminal truncation of 216-aa, or a C-terminal truncation of 132-aa with an altered C terminus

Eight full-length cDNA sequences of mouse *Cpeb3* were extracted from the UniGene database (supplementary Table 5). In addition, partial sequences of two transcripts were experimentally identified (Fig. 3A, sequences with dashed lines).[5,6] Sequence alignments indicated that the alternative usage of exons 1–3 and variable length of exon 13 lead to different 5' UTRs

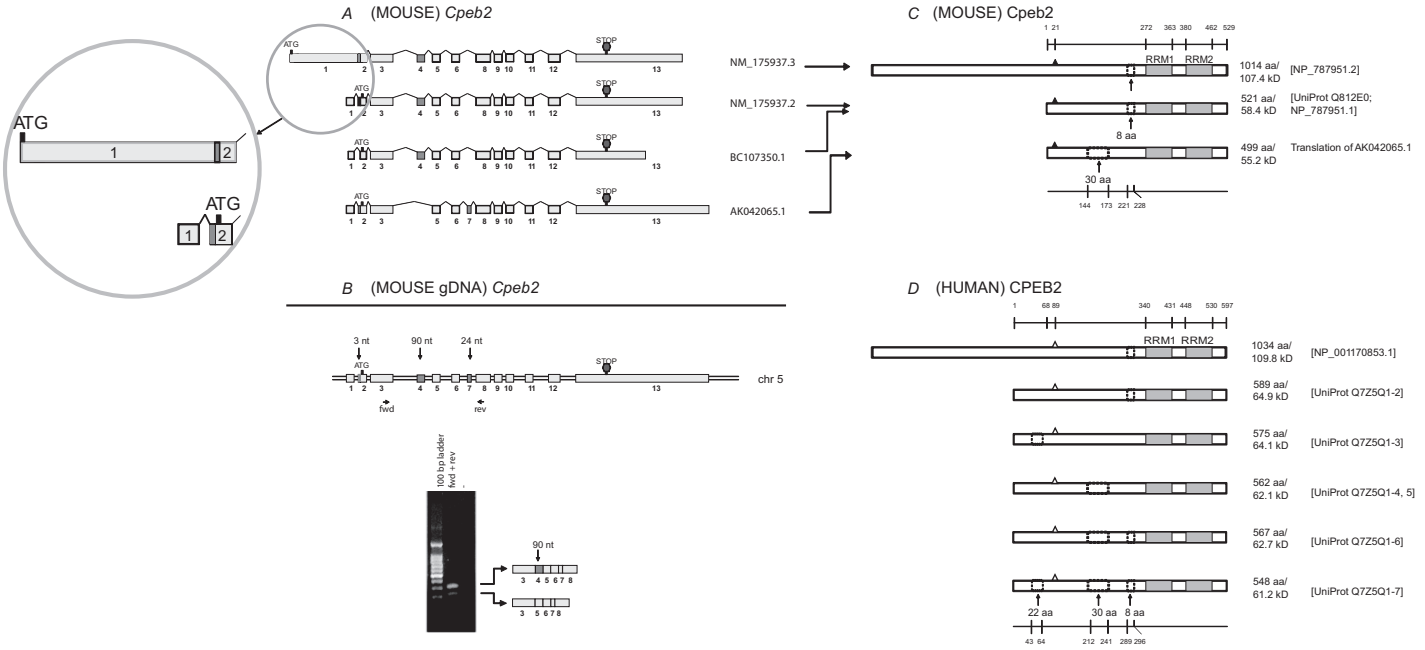


Figure 2. Analysis of *Cpeb2*. **A)** Transcripts of mouse *Cpeb2*. Three previously established non-redundant full-length UniGene sequences and a newly updated sequence (the top one) are used for the analysis, with their accession numbers listed on the right of the figure. ATG and STOP indicate the presence of translational initiation and termination sites, respectively. The different lengths of the first and the last exons likely represent the presence of variable 5' UTR and 3' UTR, respectively. The alternative splice of the first 3-nt of exon 2 also result in a different 5' UTR. The alternative splices of exon 4 (90-nt) or exon 7 (24-nt) would generate different protein products. **Insert**: the previous version and the new version of NM_175937 sequences are both legitimate. In the new version, the first exon was extended towards both ends and "fused" with exon 2. This would lead to a much longer cDNA and a longer N terminus in the protein. **B)** Expression of *Cpeb2* transcripts in adult mouse retina. The locations of the primers for RT-PCR are aligned to the diagram of *Cpeb2* genomic DNA, in which boxes represent exons and double lines represent introns. The photograph of DNA Gel demonstrates the expression of two *Cpeb2* transcripts in the retina with and without exon4. The identity of each band was confirmed with nucleotide sequencing. **C)** Isoforms of mouse Cpeb2 proteins. The previously established isoform (Q812E0; NP_787951.1) has an 8-aa deletion. The newly updated isoform (NP_787951.2) has an 8-aa deletion as well as a much longer N-terminus, which is due to the use of a longer exon 1 in cDNA NM_175937.3. The computational translation of cDNA AK042065.1 generates an additional isoform which has a 30-aa deletion. RRMs are indicated with gray boxes. Triangles represent phosphorylation sites experimentally confirmed.[22] A 30-aa deletion and an 8-aa deletion are indicated in dashed line boxes. The locations of functional motifs are shown as numbered amino acid sites at the top of the diagram, and those of the alternative spliced regions at the bottom, as might be seen in a conceptual isoform without any deletion. **D)** Isoforms of human CPEB2 proteins. Five isoforms are extracted from the UniProt database. The RRMs, the phosphorylation sites (open triangles, predicted based on cross-ortholog comparisons), and the two deletions are all present in human CPEB2 at similar locations. The recently updated, unusually long isoform of human CPEB2 (NP_001170853.1) was aligned to its closest isoform. The first 68-aa in the previously established human CPEB2 isoforms was thought to be a region that was unique for human, but now aligns to a region in the extra-long isoform of mouse Cpeb2.

or 3' UTRs, respectively (Fig. 3A). Five alternative splices within the CDS region involve intra-exon skipping of exon 4 (388-nt), partial skipping of exon 5 (69-nt), deletion of exon 7 (24-nt), deletion of exon 11 (115-nt), and extension of exon 11. The intra-exon skipping of exon 4 results in the use of an alternative translation start codon. The extension of exon 11 leads to altered downstream sequence and an early termination. The majority of these alternatively spliced regions have been experimentally identified in many tissues, including in multiple regions of the central nervous system.[6] The partial skipping of exon 5 and the deletion of exon 7 have been observed in some other vertebrates (supplementary Table 6, the asterisks).

Six protein sequences of mouse Cpeb3 were extracted from the UniProt and the NCBI protein databases. Sequence comparisons demonstrated four variable regions (Fig. 3C). The alternative usage of a 23-aa region and an 8-aa region is attributable to the alternative splicing in exon 5 and exon 7, respectively. A 216-aa N-terminal truncation may result from the use of an alternative translation initiation codon when intra-exon skipping occurs to exon 4. A 132-aa C-terminal truncation which removes the majority of the second RRM2, and terminates with four distinct amino acids (Fig. 3C, Q7TN99-5) can be derived from the extension of exon 11 (Fig. 3A).

The sequences and locations of the 23-aa and the 8-aa regions are conserved in human CPEB3 and

mouse Cpeb3 (Fig. 3C, 3D). One human isoform has a deletion of 17-aa, which includes the 8-aa and the adjacent 9-aa C-terminal to it (Fig. 3D, Q8NE35-1). To further explore the validity of the 17-aa deletion, we compared additional organisms. A particular exon (supplementary Table 6, column denoted by the pound sign) shows two variants (141-nt or 168-nt) among the species investigated. We postulated that the presence of the 141-nt was due to a 27-nt skipping within the 168-nt exon. Sequence alignment indicated that this 168-nt is exon 8 in mouse *Cpeb3* (Fig. 3A), and the 27-nt skipping would occur in the beginning of exon 8 (Fig. 3A, the part of exon 8 highlighted in gray), and would lead to a deletion in the protein product of 9-aa next to the aforementioned 8-aa. The removal of the 8-aa disrupts a Pkb recognition site,

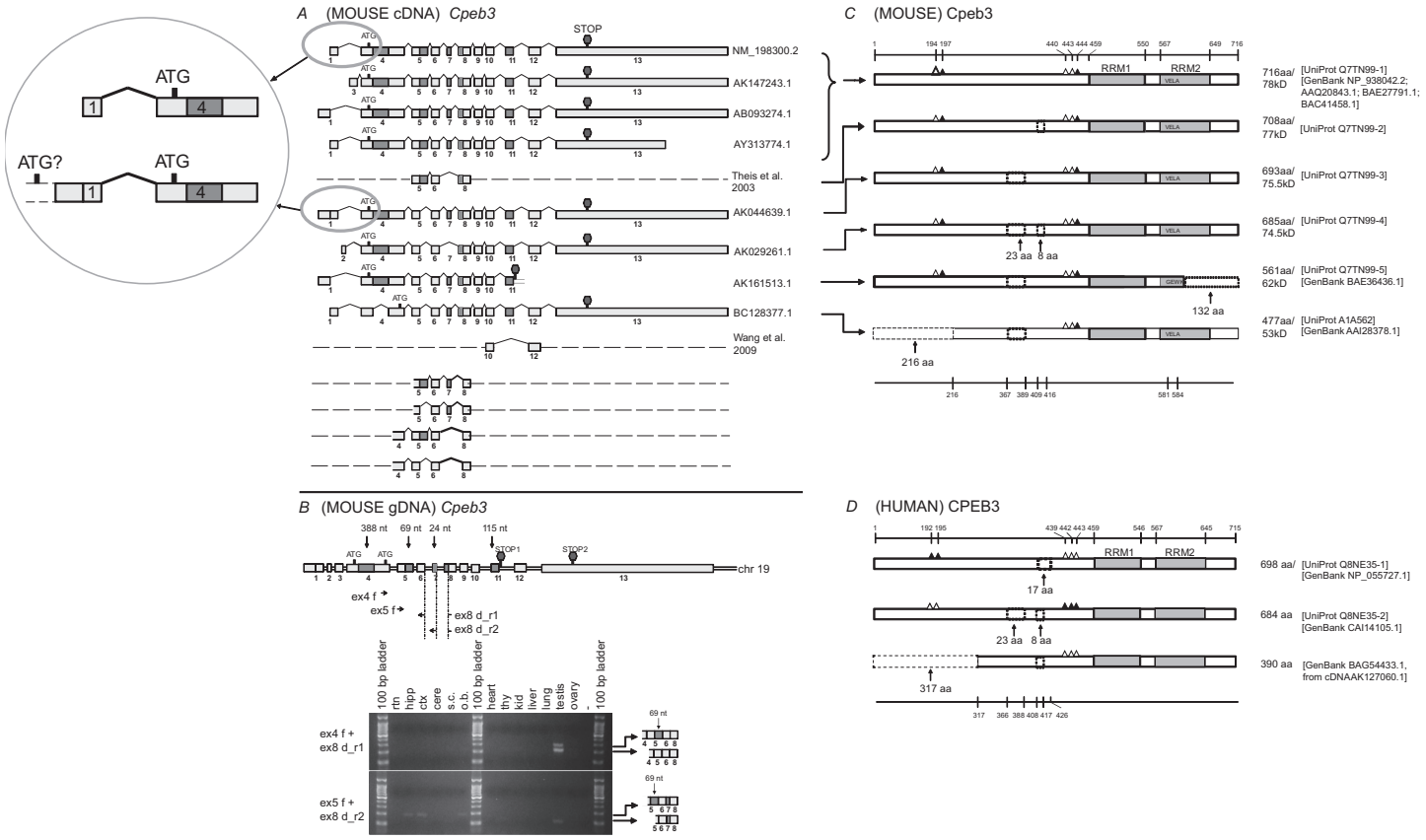


Figure 3. Analysis of *Cpeb3*. **A)** Transcripts of mouse *Cpeb3*. Eight non-redundant full-length UniGene sequences were used for the analysis, with the accession numbers listed on the right. ATG and STOP indicate the presence of translational initiation and termination sites, respectively. Partial sequences of two transcripts were derived from previous publications.[5,6] The different lengths of the first and the last exons likely represent different 5' UTR and 3' UTR, respectively. The alternative splices of exon 2 and 3 would also result in different 5' UTRs. The intra-exon skipping of exon 4 (388-nt), the partial skipping of exon 5 (69-nt), and the deletion of exon 7 (24-nt) or exon 11 (115-nt) would generate different protein products. Exon 11 extension (AK161513.1) would lead to an early translational termination and an altered 4-aa at the C-terminus. Four additional alternatively splice variants, all with a 27-nt skipping at the beginning of exon 8, were identified in this study. **Insert**: The comparison between two cDNAs that use different 5' UTRs. The upstream elongation of exon 1 in AK044639.1, with additional alternative splice(s), may lead to the use of an alternative translation start codon, which would generate an extra-long isoform of Cpeb3 protein with an extended N-terminus. **B)** Expression of *Cpeb3* transcripts in adult mouse retina. The locations of the primers for RT-PCR are aligned to the diagram of *Cpeb3* genomic DNA, in which boxes represent exons and double lines represent introns. Photographs of DNA gel demonstrate the expression of four *Cpeb3* transcripts in the retina without the 27-nt in exon 8, and with or without the 24-nt (exon 7) and the 69-nt (exon 5). The identity of each band was confirmed with nucleotide sequencing. Many of the other alternatively spliced regions have been confirmed in previous publications[5,6] in great details. Tissue abbreviation: rtn—retina, hipp—hippocampus, ctx—cortex, cere—cerebellum, s.c.—spinal cord, o.b.—olfactory bulb, thy—thymus, kid—kidney. **C)** Isoforms of mouse Cpeb3 proteins. Six isoforms are extracted from the UniProt database. RRMs were indicated with gray boxes. Triangles represent phosphorylation sites experimentally confirmed[21,24,25] (solid) or predicted (open) according to cross-paralog comparisons, respectively. A 216-aa N-terminal truncation, two internal deletions of 23-aa motif and 8-aa motif, and a 132-aa C terminal truncation with altered C terminus were indicated in dashed lines. The C-terminal truncation (Q7TN99-5) removes the majority of the second RRM and alters the last four amino acids from VELA to GEWK. The locations of functional motifs are shown as numbered amino acid sites at the top of the diagram, and those of the alternative spliced regions at the bottom, as might be seen in the longest isoform. **D)** Isoforms of human CPEB3 proteins. Three isoforms are extracted from UniProt and NCBI databases. The RRMs, the phosphorylation sites, and the 23-aa and 8-aa deletions are all present in human CPEB3 at similar locations. Numeric annotations refer to a conceptual isoform without any deletion.

whereas the deletion of the 17-aa abolishes the Pkb phosphorylation site as well as a Pka phosphorylation site (Table 2). The 27-nt deletion has been detected in mouse testis and to a lesser degree in the hippocampus, cortex, and olfactory bulb (Fig. 3B).

One human CPEB3 isoform (BAG54433.1) has a 317-aa N-terminal truncation (Fig. 3D). The strong homology between mouse and human prompted us to question the validities of the 317-aa truncation in the human and the 216-aa truncation in the mouse (Fig. 3C). The 216-aa truncation in mouse Cpeb3 is derived from an intra-exon skipping of exon 4 (Fig. 3A). The corresponding cDNA (AK127060.1) for human CPEB3 isoform with 317-aa truncation has a very short 5' UTR. Computational translation indicated that the truncation of 317-aa could be due to an incomplete 5' sequence of the cDNA. Based on the stringent conservation between mouse and human, particularly in the alternatively spliced regions (Fig. 6 C–D), it is possible that this human CPEB3 protein isoform has an N terminal truncation of 216-aa instead of 317-aa. Techniques such as 5' rapid amplification of cDNA ends (5' RACE) may be used in future studies to obtain the complete 5' sequence of this transcript as a means to confirm the length of N-terminal truncation in human CPEB3.

More drastic changes appear in the Cpeb3 isoform with a 132-aa C-terminal truncation and an altered tail (VELA→GEWK) (Fig. 3C, and yellow box in Fig. 5). This isoform lacks the majority of the second RRM, including its octamer consensus. This is likely to have a significant impact on RNA-protein interaction and specificity. The DNA or RNA binding proteins thus far identified have one to four RRMs.[16] NMR characterization of the structures of several proteins revealed different binding mechanisms for even- and odd-numbered RRMs. For example, heterogeneous nuclear ribonucleoprotein A1 (Hnrnpa1) and polypyrimidine tract binding protein (Ptbb), which contain two and four RRMs, respectively, form homodimers when binding to two molecules of mRNA in anti-parallel arrangement.[17,18] In contrast, poly (A) binding protein 2 (Pabp2) which has a single RRM, forms homodimers in the absence of RNA, but becomes monomeric upon mRNA binding.[19] Thus the Cpeb3 isoform with a 132-aa C-terminal truncation and an altered tail would likely have its binding characteristics altered.

The largest predicted size for mouse Cpeb3 is approximately 78 kD, but a protein greater than 100 kD has also been detected with antibodies in western blots.[6,8] This larger protein has been proposed to be a pre-protein.[6] Since Cpeb2 and Cpeb3 are closely-related in the Cpeb family (Fig. 6A), the recent finding of the extra-long Cpeb2 (Fig. 2A, C, D, the top isoforms) makes it plausible to postulate the presence of a similar extra-long isoform for Cpeb3. Both human and chimpanzee have a beginning exon of 193-nt (supplementary Table 6, the double pound signs) which, if mapped to mouse genome, would be adjacent to the 61-nt exon. In addition, one isoform of mouse *Cpeb3* indeed has an extended "5' UTR" (Fig. 3A, AK044639.1). A putative translation indicates that an upstream extension of the 61-nt exon into and beyond the 193-nt would lead to a continuous extension of Cpeb3 protein beyond the N-terminus. We analyzed a genomic sequence of about 2000-bp upstream of the first exon, and realized that for the extended translation to be long enough (that is, to match the difference between ~100 kD and 78 kD), additional upstream splice(s) may be necessary (Fig. 3, insert). Additional experimental evidence is required to determine the validity and the exact length of an extra-long Cpeb3.

Cpeb4 protein isoforms with internal deletions of 17-aa or 8-aa, or an N-terminal truncation of 382-aa

Sixteen cDNA sequences representing mouse *Cpeb4* were extracted from the UniGene database (supplementary Table 7). After removing the fragmented and redundant sequences, we used three remaining cDNA sequences for sequence alignment. Two additional isoforms based on a previous report were also used for analysis[5] (Fig. 4A). The comparison of these five sequences demonstrated that variations in exon 1 could lead to different 5' UTRs or alternative translation initiation sites. Variations in the length of the last exon could lead to different 3' UTRs. Alternative splicing of exon 3 and exon 4 would likely result in the removal of 51-nt and 24-nt, respectively (Fig. 4A). All four isoforms related to altered exons 3 and 4 have been identified in mouse brain tissue.[5] Two isoforms in the same region are also evident in adult mouse retina (Fig. 4B). The alternative use of

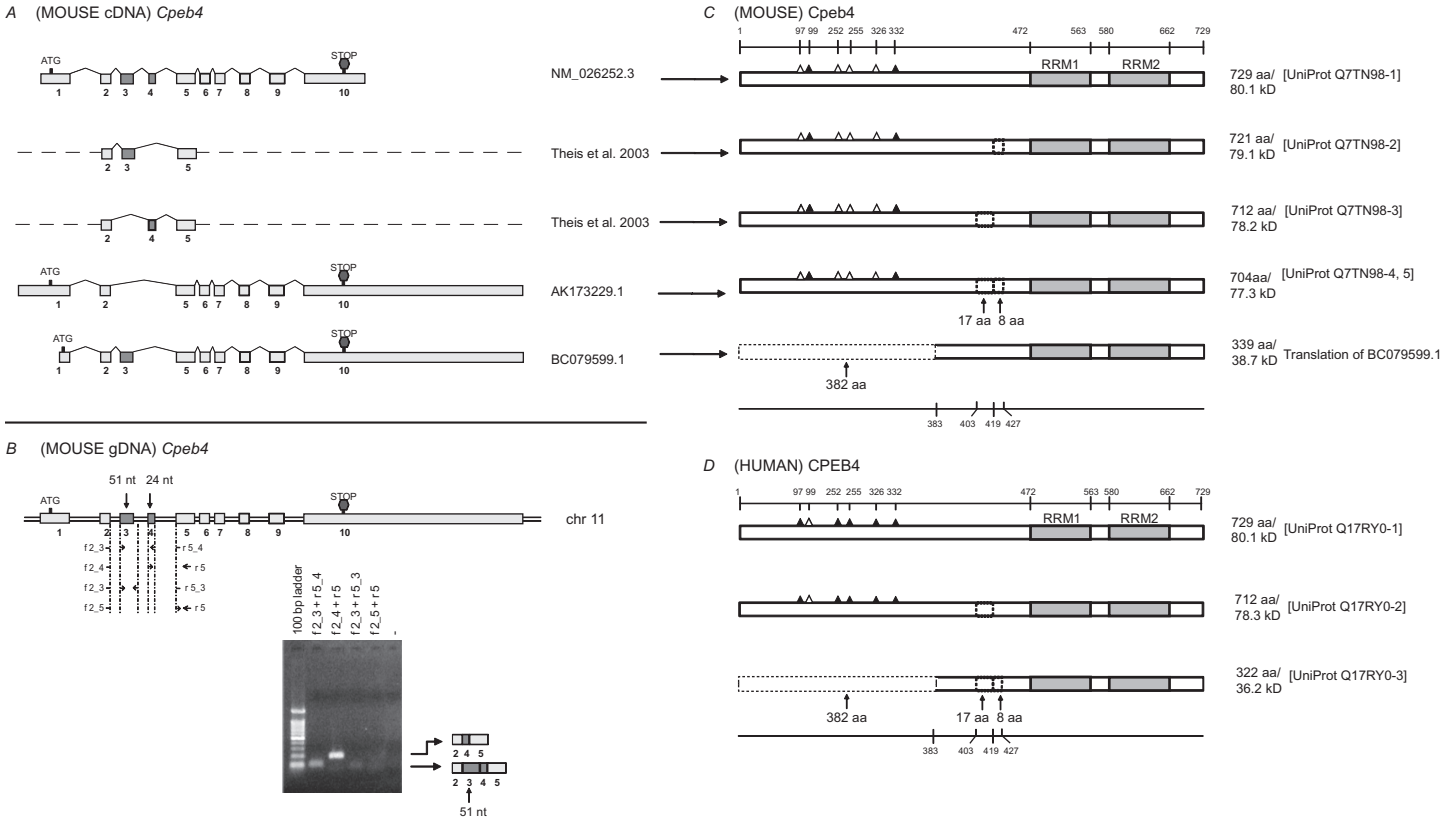


Figure 4. Analysis of *Cpeb4*. **A)** Transcripts of mouse *Cpeb4*. Three non-redundant full-length UniGene sequences were used for this analysis, with the accession numbers listed on the right. ATG and STOP indicate the presence of translational initiation and termination sites, respectively. Partial sequences of two transcripts are derived from a previous publication.[5] The different lengths of the first and the last exons likely represent the presence of different 5' UTR and 3' UTR, respectively. The alternative splices of exon 3 (51-nt) or exon 4 (24-nt) would generate different protein products. A shorter first exon leads to an alternative translational initiation site in BC079599.1. **B)** Expression of *Cpeb4* transcripts in the adult mouse retina. The locations of the primers for RT-PCR are aligned to the diagram of *Cpeb4* genomic DNA, in which boxes represent exons and double lines represent introns. The photograph of DNA Gel demonstrates the expression of multiple *Cpeb4* transcripts in the retina with and without the exon 3. The identity of each band was confirmed by nucleotide sequencing. **C)** Isoforms of mouse Cpeb4 proteins. Four isoforms were extracted from the UniProt database. The computational translation of cDNA BC079599.1 would generate an additional isoform. RRMs are indicated with gray boxes. Triangles represent phosphorylation sites experimentally confirmed (solid)[21,23] or predicted (open) based on cross-paralog comparisons, respectively. A 17-aa deletion, an 8-aa deletion, and a 382-aa N-terminal truncation are each indicated with dashed line boxes. The locations of functional motifs are shown as numbered amino acid sites at the top, and those of the alternative spliced regions at the bottom, as might be seen in the longest isoform. **D)** Isoforms of human CPEB4 proteins. Three isoforms are extracted from the UniProt database. The RRMs, the phosphorylation sites, the deletions and the truncation are all present in human CPEB4 at the same locations. All numerical locations refer to the longest isoform.

these two exons in *Cpeb4* is highly conserved among vertebrates (supplementary Table 8, the asterisks).

Four isoforms of mouse Cpeb4 proteins were extracted from the UniProt database. The alignment of the protein isoforms reflects the deletions of 17-aa and 8-aa motifs (Fig. 4C), which correspond to the removal of exon 3 and exon 4, respectively (Fig. 4A). Computational translation of cDNA BC079599.1 yielded a novel protein with a 382-aa N-terminal truncation (Fig. 4C). The deletion of 382-aa, like the deletions of the 17-aa, and 8-aa, was identified in human CPEB4 at the same locations (Fig. 4D). These sequences of the alternatively spliced regions were also strictly conserved in mouse and human (Fig. 6D).

Across-paralog comparison of mouse Cpeb1, 2, 3 and 4

The overall sequence of *Cpeb1* has been reported to have low homology to *Cpeb2–4*.[5] We demonstrate here that the alternatively spliced regions of Cpeb1 are rather different from those of Cpeb2–4 (Fig. 5, 6C–D). This fact strengthens the previously reported notion that Cpeb1 is a distant cousin of Cpeb2–4. Cpeb2, 3, and 4 have almost identical RRMs but variable N termini. Of interest, an 8-aa motif within the variable region is stringently conserved among Cpeb2–4. This motif is located N-terminal to the first RRM (Fig. 5, the red box). Its deletion leads to the removal of certain functional motifs for protein

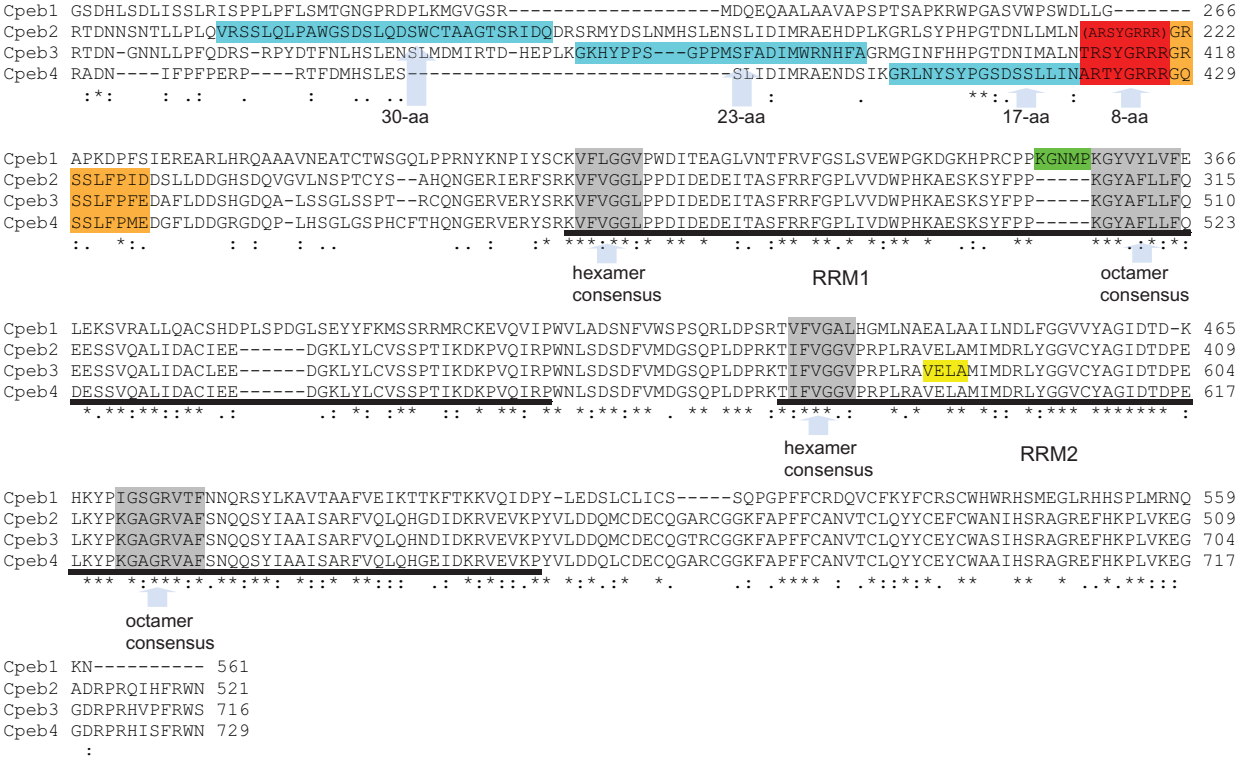


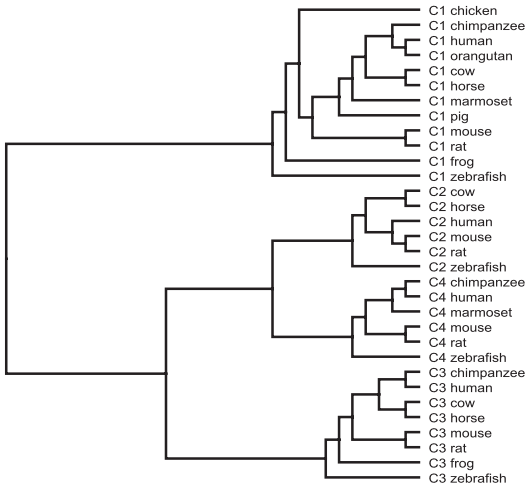
Figure 5. Comparison of the conserved regions in mouse Cpeb1–4 protein. The alternatively spliced 17~30-aa regions are highlighted in blue, the alternatively spliced 8-aa in red, and the alternatively spliced 9-aa in orange. The parentheses indicate that all Cpeb2 isoforms recorded in the UniProt database have the 8-aa deleted. Cpeb2, 3, and 4 share high homology in the 8-aa and the 9-aa regions, but little homology in the 17~30-aa region. The underlined sequences represent the first and second RRMs, respectively, in all four CPEBs. The regions in grey are hexamer and octamer consensus sequences within the RRMs. The hexamer and octamer consensus sequences within the RRMs and the linker between two RRMs are identical in Cpeb2–4, suggesting that it is highly likely that Cpeb2–4 share the same protein/RNA interaction mechanisms. The sequences surrounding the consensuses, N terminal to the first RRM, and C terminal to the second RRM are similar among Cpeb2–4, with a few amino acid replacements. This suggests that Cpeb2–4 recognizes similar substrates. In contrast, Cpeb1 demonstrates significant differences to Cpeb2–4 within these regions, including the consensus sequences, suggesting that Cpeb1 not only employs a distinct mechanism for protein/RNA interaction, but also targets a different group of RNAs. The insertion of the 5-aa in Cpeb1 (highlighted in green) is adjacent to the octamer consensus in the first RRM, possibly posing a potential impact on its specificity. An early termination with an altered tail which alters VELA (highlighted in yellow) to GEWK disrupts the second RRM in a Cpeb3 isoform. This may pose an impact on both the binding mechanism and the substrate specificity of Cpeb3. Accession numbers used for the alignment are as follows: Cpeb1: NP_031781, Cpeb2: NP_787951; Cpeb3: NP_938042; Cpeb4: NP_080528. Alignment was achieved with the aid of ClustalW2. Asterisks represent perfect matches; colons represent substitutions with similar amino acids; periods represent substitutions with rather distinct amino acids.

cleavage, protein-protein interaction, and phosphorylation (Table 1–3). Although the deletion of the 8-aa disrupts a Pkb recognition site, the newly identified 9-aa deletion adjacent to the 8-aa in Cpeb3 (Fig. 5, orange box) would lead to the removal of this Pkb phosphorylation site and a Pka phosphorylation site altogether (Table 2). Based on the sequence homology among Cpeb2–4, it is plausible to predict that the 9-aa may be alternatively spliced in Cpeb2 and Cpeb4 as well. The deletion of the 9-aa would likely cause the removal of a Pka phosphorylation site in Cpeb2 as well (Table 1). However, this Pka phosphorylation site is absent from Cpeb4 due to a single amino acid substitution (GRSSLLP -> GQSSLLP, Fig. 5).

Another common feature of Cpeb2–4 is the alternatively splicing of the 17~30-aa N-terminal to the 8-aa motif. In contrast to the 8-aa and 9-aa sequences which are highly conserved among Cpeb2–4, the 17~30-aa sequences show little evidence of homology among the three paralogs (Fig. 5, the blue boxes). Functional predictions demonstrated that the deletion of this region removes motifs implicated in protein-protein interactions, phosphorylation, and post-translational modifications (Table 1–3). Noteworthy among these findings is the deletion of the 23-aa motif in Cpeb3: this not only removes certain functional motifs, but also creates a novel site for Mapk interaction (Table 2).

This 17~30-aa variable regions become shorter and closer to the 8-aa motif in the order of Cpeb2→Cpeb3→Cpeb4, until the gap closes in Cpeb4 (Fig. 5). Functional predictions revealed that the linker regions

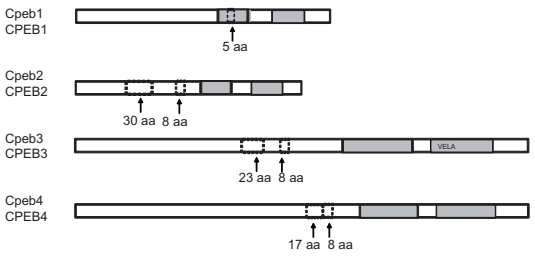
A



B

Orthologs	Homology	Comments
mouse Cpeb1 vs. human CPEB1	95%	Both canonical forms. Exclude an extra 4-aa at human C-terminus.
mouse Cpeb2 vs. human CPEB2	99%	Both with 8-aa deletion. Exclude an extra 68-aa at human N-terminus.
mouse Cpeb3 vs. human CPEB3	96%	Both with 8-aa deletion.
mouse Cpeb4 vs. human CPEB4	98%	Both canonical forms.

C



D

1-aa Cpeb1- A → A absent in human

5-aa Cpeb1- (DGKHPRCPPK) GNMPK (GYVYLVFELEK)

17~30-aa region
Cpeb2- (NNSNTLLPLQ) VRSSLQLPAWGSDSLQDSWCTAAGTSRIDQ
Cpeb3- (IRTDHEPLKG) GKHYPPSGPPMSFADIMWRNHFA
Cpeb4- (IMRAENDSIK) GRLNYSYPGSDSSLLIN → S to T in human

Linker
Cpeb2- DRSRMYDSLNMHSLENSLIDIMRAEHDPLKGRLSYPHPGTDNLLMLN
Cpeb3- GRMGINFHHPGTDNIMALN
Cpeb4- none

8-aa region
Cpeb2- ARSYGRRR (GRSSLFPIDD)
Cpeb3- TRSYGRRR(GRSSLFPFED) → G to S in human
Cpeb4- ARTYGRRR (GQSSLFPMED)

Figure 6. Comparisons between (MOUSE) Cpebs and (HUMAN) CPEBs. **A)** The NM_sequences from multiple species (complete sequences) are used for the generation of a phylogenetic tree. The tree demonstrated that the distances between orthologs are significantly closer than those between paralogs. **B)** Percentage homology between mouse and human for each CPEB protein. RefSeq (NM_) sequences are used for the comparison. Certain regions in human proteins are excluded from the estimate of percentage homology as indicated in the comments. **C)** The patterns and locations of deletions have little or no difference in mouse and human. In Cpeb1, the 5-aa alternative splice occurs within the first RRM. In Cpeb2–4, the 8-aa deletion is always located N-terminal to the RRMs, and the 17~30-aa deletion N-terminal to the 8-aa sequence. The 17~30-aa sequence becomes shorter and closer to the 8-aa in the order of Cpeb2→Cpeb3→ Cpeb4, until the gap closes in Cpeb4. The same is true for human CPEBs. **D)** Sequence comparisons of the variable regions between mouse Cpebs and human CPEBs. The majority of the alternatively spliced sequences and the adjacent sequences (indicated in parentheses) are identical in mouse and human, with only a few single nucleotide substitutions (in red).

may harbor class III PDZ (postsynaptic density-95, discs-large, zonula occludens-1) domain binding motifs. The numbers of such motifs vary: there are 3 in CPEB2 (Table 1), 1 in Cpeb3 (Table 3), and none in Cpeb4 (Table 5). The sequences of class III PDZ domain binding motifs are also different. Such motifs may recruit different Cpeb paralogs to different protein partners, thereby leading to distinctive localizations and functions.

Together, the aforementioned variations enclosing the 8-aa, 9-aa, and 17~30-aa region would likely determine protein-protein interaction, phosphorylation, and post-translational modifications of Cpebs. The functional significance of this region is of great interest for future studies.

Three paralogs, Cpeb1, 3 and 4, have isoforms with large N-terminal truncations. Such truncations may have a major impact on the function of the proteins. For instance, the Cpeb4 isoform with a large N terminal truncation may be deprived of many, if not all, of the phosphorylation sites (Fig. 4C). This would likely make it a putative candidate for a dominant negative form. Our analysis using the bioinformatics tool Eukaryotic Linear Motif resource (ELM, http://elm.eu.org) indicated that the N-terminal fragments may also harbor featured sites such as those required for post-translational modification and protein-protein interaction (data not shown). The presence or absence of such sites may alter signaling pathways, stimulus-dependence, or the development of particular protein complexes.

The C termini of RNA binding proteins determine the affinity and specificity of RNA binding. Two highly conserved short consensus motifs, a hexamer and an octamer, are separated by about 30-aa and embedded in a structurally conserved, but not sequence conserved RRM region of approximately 90-aa[16,20] (Fig. 5). These two short consensuses are deemed hallmarks of RNA binding proteins. The linker sequences between two RRMs are also highly conserved among RNA binding proteins. The hexamer and octamer consensuses and the linker regions are essential for protein-RNA interaction. However, the specificity of RNA binding is determined by sequences surrounding the hexamer and octamer, as well as sequences N terminal to the first RRM and C terminal to the second RRM.[14,15] Based on the near-identical homology of these important functional regions in Cpeb2, 3, and 4 (Fig. 5),

Table 1. Cpeb2 motifs spanning the 30-aa and 8-aa predicted with the aid of ELM. The motifs in red are missing in the isoform with 8-aa deletion; the ones in dark blue are missing in the isoform with 30-aa deletion; the ones in both colors are missing in the isoform with 30-aa and 8-aa deletions. The motifs in green are located in the linker region. The motif in light blue is missing in the newly identified 9-aa deletion.

Elm name	Instances	Positions	Elm description	Pattern
CLV_NDR_NDR_1	RRG	94–96	N-Arg dibasic convertase (nardilysine) cleavage site (Xaa-|-Arg-Lys or Arg-|-Arg-Xaa)	.RK|RR[^KR]
CLV_PCSK_PC7_1	RSYGRRR	89–95	Proprotein convertase 7 (PC7, PCSK7) cleavage site (Arg-Xaa-Xaa-Xaa-[Arg/Lys]-Arg-|-Xaa)	[R]...[KR]R.
LIG_14-3-3_2	RMYDSLN	44–50	Longer mode 2 interacting phospho-motif for 14-3-3 proteins with key conservation RxxxS#p	R..[^P][ST][IVLM].
	RGRSSLF	95–101	Longer mode 2 interacting phospho-motif for 14-3-3 proteins with key conservation RxxxS#p.	
LIG_APCC_Dbox_1	GRSSLFPID	96–104	An RxxL-based motif that binds to the Cdh1 and Cdc20 components of APC/C thereby targeting the protein for destruction in a cell cycle dependent manner	.R..L..[LIVM].
LIG_BRCT_BRCA1_1	RSSLF	97–101	Phosphopeptide motif which directly interacts with the BRCT (carboxy-terminal) domain of the Breast Cancer Gene BRCA1 with low affinity	.S..F
LIG_Clathr_ClatBox_1	LFPID	100–104	Clathrin box motif found on cargo adaptor proteins, it interacts with the beta propeller structure located at the N-terminus of Clathrin heavy chain.	L[IVLMF].[IVLMF][DE]
LIG_FHA_1	AGTSRID	33–39	Phosphothreonine motif binding a subset of FHA domains that show a preference for a large aliphatic amino acid at the pT + 3 position.	..(T)..[ILV].
	PGTDNLL	78–84		
LIG_PDZ_3	SDSL	22–25	Class III PDZ domains binding motif	.[DE].[IVL]
	YDSL	46–49	Class III PDZ domains binding motif	
	HDPL	66–69		
	TDNL	80–83		
LIG_USP7_1	AWGSD	19–23	The USP7 NTD domain binding motif variant based on the MDM2 and P53 interactions.	[PA][^P][^FYWIL]S[^P]
	AGTSR	33–37		
MOD_CK1_1	SWCTAAG	28–34	CK1 phosphorylation site	S..([ST])...
MOD_Cter_Amidation	YGRR	91–94	Peptide C-terminal amidation	(.)G[RK][RK]
MOD_GSK3_1	GSDSLQDS	21–28	GSK3 phosphorylation recognition site	...([ST])...[ST]
	SWCTAAGT	28–35		
	NMHSLENS	50–57		
MOD_PKA_2	VRSSLQL	11–17	Pka phosphorylation site	.R.([ST])...
	GRLSYPH	71–77		
	GRSSLFP	96–102	Pka phosphorylation site	
MOD_PKB_1	RRRGRSSLF	93–101	Pkb Phosphorylation site	R.R..([ST])...
MOD_PLK	LENSLID	54–60	Site phosphorylated by the Polo-like-kinase	.[DE].[ST][ILFWMVA]..
TRG_ENDOCYTIC_2	YDSL	46–49	Tyrosine-based sorting signal responsible for the interaction with mu subunit of AP (Adaptor Protein) complex	Y..[LMVIF]

Note: The numeric positions in column 3 are based on the following sequence investigated: NNSNTLLPLQ**VRSSLQLPAWGSDSLQDSWCTAAGTSRIDQ**DRSRMYDSLNMHSLENSLIDIM RAEHDPLKGRLSYPHPGTDNLLMLN**ARSYGRRR**GRSSLFPIDD. The letters in bold represent the 30-aa and the 8-aa regions.

Table 2. The motifs in Cpeb3 spanning the 23-aa and 8-aa regions predicted with the aid of ELM. The motifs in red are missing in the 8-aa deletion isoform; the ones in dark blue are missing in the 23-aa deletion isoform; the ones in both colors are missing in the isoform with 23-aa and 8-aa deletions. The motif in pink appears only in the isoforms with 23-aa deletion, or 23-aa and 8-aa deletion. The motifs in green are located in the linker region. The motif in light blue is missing in the newly identified 9-aa deletion.

Elm name	Instances	Positions	Elm description	Pattern
CLV_NDR_NDR_1	RRG	58–60	N-Arg dibasic convertase (nardilysine) cleavage site (Xaa-|-Arg-Lys or Arg-|-Arg-Xaa)	.RK|RR[^KR]
CLV_PCSK_PC7_1	RSYGRRR	53–59	Proprotein convertase 7 (PC7, PCSK7) cleavage site (Arg-Xaa-Xaa-Xaa-[Arg/Lys]-Arg-|-Xaa)	[R]...[KR]R.
LIG_14-3-3_2	RGRSSLF	59–65	Longer mode 2 interacting phospho-motif for 14-3-3 proteins with key conservation RxxxS#p.	R..[^P][ST][IVLM].
LIG_BRCT_BRCA1_1	RSSLF	61–65	Phosphopeptide motif which directly interacts with the BRCT (carboxy-terminal) domain of the Breast Cancer Gene BRCA1 with low affinity	.S..F
LIG_Clathr_ClatBox_1	LFPFE	64–68	Clathrin box motif found on cargo adaptor proteins, it interacts with the beta propeller structure located at the N-terminus of Clathrin heavy chain.	L[IVLMF].[IVLMF][DE]
LIG_EVH1_II	PPMSF	18–22	Proline-rich motif binding to signal transduction class II EVH1 domains	PP..F
LIG_FHA_1	PGTDNIM	42–48	Phosphothreonine motif binding a subset of FHA domains that show a preference for a large aliphatic amino acid at the pT + 3 position.	..(T)..[ILV].
LIG_FHA_2	IRTDHEP	1–7	Phosphothreonine motif binding a subset of FHA domains that have a preference for an acidic amino acid at the pT + 3 position.	..(T)..[DE].
LIG_PDZ_3	HEPL TDNI	5–8 44–47	Class III PDZ domains binding motif	.[DE].[IVL]
LIG_SH3_3	HYPPSGP	12–18	This is the motif recognized by those SH3 domains with a non-canonical class I recognition specificity	...[PV]..P
MOD_Cter_Amidation	YGRR	55–58	Peptide C-terminal amidation	(.)G[RK][RK]
MOD_GlcNHglycan	PSGP	15–18	Glycosaminoglycan attachment site	[ED](0,3).(S)[GA].
MOD_PKA_2	GRSSLFP	60–66	Pka phosphorylation site	.R.([ST])...
MOD_PKB_1	RRRGRSSLF	57–65	Pkb Phosphorylation site	R.R..([ST])...
TRG_PEX	WRNHF	27–31	Specific ELM present in Pex5p and binding to Pex13p and Pex14p. Part of the peroxisomal matrix protein import system	W...[FY]
LIG_MAPK_1	KGRMGINF	9–16 (Number refer to Sequence* and **).	MAPK interacting molecules (e.g. MAPKKs, substrates, phosphatases) carry docking motif that help to regulate specific interaction in the MAPK cascade. The classic motif approximates (R/K)xxxx#x# where # is a hydrophobic residue.	[KR](0,2)[KR].(0,2)[KR].(2,4)[ILVM].[ILVF]

Note: The numeric positions in column 3 refer to positions in the following sequence investigated: IRTDHEPLK**GKHYPPSGPPMSFADIMWRNHFA**GRMGINFHHPGTDNIMALN**TRSYGRRR**GRSSLFPFED with the exception that the number in the last row refers to positions in the following two sequences: *IRTDHEPLKGGRMGINFHHPG TDNIMALN**TRSYGRRR**GRSSLFPFED and **IRTDHEPLKGGRMGINFHHPGTDNIMALNGRSSLFPFED. The letters in bold represent the 23-aa and/or the 8-aa regions.

we predict that Cpeb2, 3, and 4 recognize similar targets. However, Cpeb1, whose sequence deviates significantly from that of Cpeb2–4 even with regard to the short consensus, must recognize a different set of targets and employ a distinct mechanism for RNA interaction. Indeed certain RNA oligonucleotides interact with Cpeb3 and Cpeb4, but not Cpeb1 protein, and the reverse is also true.[8]

Across-ortholog comparisons provide new insights

Comparisons across species can be instructive. In this study we found that the exon structures of *Cpeb* orthologs among vertebrates are almost identical. Most of the internal exons are of exactly the same lengths across a wide range of organisms from zebrafish to human (supplementary tables 2, 4, 6, 8). With regard to the proteins, the phylogenetic tree clearly demonstrated that they are better conserved across species than across paralogs (Fig. 6A). Of all currently documented Cpeb proteins, the orthologs are closer to each other than the paralogs are. It is also evident that Cpeb protein paralogs are highly conserved between mouse and human. Despite an additional 4-aa C-terminal "tag" in human CPEB1 (Fig. 1D), the rest of the sequences between mouse and human orthologs are nearly identical (Fig. 6B). The patterns, locations, and sequences of the alternatively spliced regions are strictly preserved between mouse and human (Fig. 6C, 6D). For instance, the 5-aa deletion in mouse Cpeb1, although not present in mouse Cpeb2–4, is found to be identical in human CPEB1. This variation is also evident at the level of the transcripts in multiple species (supplementary Table 2). Similar findings are true in the alternative spliced regions in Cpeb2–4, as demonstrated for multiple vertebrates at the level of the transcripts (supplementary tables 4, 6, 8), with little or no variation between mouse and human at the level of the proteins (Fig. 6D). These findings provide strong foundations for cross-species predictions. For example, novel isoforms in one species may be predicted based on the evidence in another, as demonstrated by the discovery of the "exon 2 deletion" transcript of mouse *Cpeb1* (Fig. 1A, 1C, 1D), and the partial skipping of exon 8 in mouse *Cpeb3* (supplementary Table 6, Figure 3A, 3D). Information in one species may also be borrowed to cross-examine the accuracy in another, as in the question regarding the length of N-terminal truncation of human CPEB3. Evidently, such comparative analysis may be used to discover unknown isoforms or to establish the correct sequences for known isoforms.

One may also exploit such logic to predict functional motifs in one species according to the other. A good example is the prediction of phosphorylation sites (PhophoSitePlus). A few amino acids have been confirmed as phosphorylation sites in mouse or human (Figs. 1–4C, D, the solid triangles) by various techniques including nuclear magnetic resonance (NMR) and mass spectrometry (MS).[21,25] Based on the stringent homology between mouse and human, the same amino acids at identical or similar locations were identified and predicted to be phosphorylation sites in the other species (Figs. 1–4C, D, the open triangles).[31]

Multiple levels of variability indicate extraordinary complexity in the regulation and function of *Cpebs*

The presence of more than one isoform in each *Cpeb* reveals the complexity in their regulatory capabilities and functions. Each alternative splicing may confer an additional layer of divergence in the regulation of biological and cellular functions. Variances in the UTRs, in particular, may attest to regulations at the transcriptional level, that is, alternative transcription initiation or termination. Variations in the UTRs then impose additional controls over translation, specifically, the initiation, termination, and efficiency of translation. Another layer of regulation, alternative splicing, leads to variances in the protein sequences themselves. The differences in protein sequences dictate the uniqueness of their functions. Alterations of as small as a few amino acids (for example, the 5-aa insertion in Cpeb1, the 8-aa and the 9-aa deletions in Cpeb3) may alter the when, where and how the Cpebs perform their functions and connect with their targets.

Conclusions

In conclusion, our study delineated alternative splicing isoforms for mouse *Cpeb1–4*. New isoforms were predicted based on theoretical translation, cross-ortholog comparison, and experimental validation. Functions of the alternatively spliced regions were predicted using bioinformatics approaches. The variety

of transcript structures and protein structures indicate an extraordinary complexity in the regulation and functions of the Cpebs.

Methods

Animal handling and tissue collection

All animal experimental procedures were performed in compliance with animal care regulations set by the University of Louisville Institutional Animal Care and Use Committee (IACUC) as well as the Association for Research in Vision and Ophthalmology (ARVO) statement for the use of animals in vision research.

C57/BL6 mice (Charles River Laboratories, Davis, CA) were used in this study to confirm the presence and/or absence of some isoforms predicted by *in silico* methods. The animals were euthanized with CO_2 followed by cervical dislocation. Retinas were dissected immediately and frozen on dry ice before proceeding to RNA extraction.

RNA extracton and RT-PCR

Frozen retinas were homogenized using a PowerGen 250 homogenizer (Fisher Scientific, Pittsburgh, PA). Total RNA was extracted using RNeasy mini kits (Qiagen, Valencia, CA) following the manufacturer's instructions. The concentration of RNA was determined using a BioPhotometer (Eppendorf, Westbury, NY), and the quality of RNA was determined by the ratio of 28S/18S on an agarose gel. RNA was frozen in −80 °C for long term storage.

0.2 µg of total RNA was used in a 20 µl RT reaction using AMV reverse transcriptase (Promega, Madison WI). 1 µl of the cDNA was used for subsequent PCR. The gene-specific primers spanning the regions of interest for *cpeb1*, *cpeb2*, *cpeb3*, and *cpeb4* were designed using Vector NTI (Analysis module, Oligo Analysis, Invitrogen, Carlsbad, CA) and obtained from IDT (Coralville, IA). Locations of these primers were demonstrated in Figure 1–4B. Sequences of these primers were listed in supplementary Table 9. PCR was carried out on a thermocycler using the following conditions: 95 °C 15 min for the initial activation; 40 cycles of: 94 °C 30 sec (denaturation), 50–55 °C 30 sec (annealing temperature varies based on the primers), 72 °C 30 sec (extension); and followed by 72 °C 10 min (final extension). The resulting PCR products were separated on 1% agarose gels and photographed. Individual bands were excised, purified and sequenced to confirm their identities.

Gene nomenclature

All gene symbols in the manuscript abide by the guidelines recommended by Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC), and are in accordance with the human HGNC database and the mouse genome databases (MGD). For example, *Cpeb* represents mouse DNA or mRNA, Cpeb represents mouse protein, *CPEB* represents human DNA or mRNA, and CPEB represents human protein.

Data mining, sequence alignment, and theoretical translation

Both mouse curated RefSeq sequences (NM_) and uncurated cDNAs were extracted from the UniGene database (www.ncbi.nlm.nih.gov/unigene) to collect as much information as possible. For all the other species, only RefSeq sequences were extracted for simplicity. The genomic sequences were derived from UCSC genome database (www.genome.ucsc.edu). UCSC Blat (www.genome.ucsc.edu/cgi-bin/hgBlat; Genome: mouse; at default settings) was used to align cDNA sequences to the genome, to deduce the length of exons, and to define the boundaries of exons. The location of each alternative splice was determined by comparing the genomic locations of exon-exon boundaries. NCBI Blast was used to align cDNAs to one another and to remove redundant sequences from further analysis. Whenever possible, partial sequences encompassing alternative regions of the cDNA entries were confirmed in our laboratory using RT-PCR and subsequent sequencing.

Mouse and human protein sequences were extracted from the UniProt database (www.uniprot.org/). The NCBI protein database (www.ncbi.nlm.nih.gov/protein/) was also explored for additional information. Mouse protein sequences were compared to mouse cDNA sequences with the aid of computational translation using Vector NTI software (Analyses module, Translation). Six frames (3 direct, 3 complementary) were used for each translation. The frame that gave the longest continuous read was selected, and the product designated as the protein product. If the translation started with the first codon which is not a methionine, or terminated at the last codon which is not a stop codon, then the cDNA is

Table 3. Cpeb4 motifs spanning the 17-aa and 8-aa regions predicted with the aid of ELM. The motifs in red are missing in the 8-aa deletion isoform; the ones in dark blue are missing in the 17-aa deletion isoform; motifs in both colors are missing in the isoform with 17-aa and 8-aa deletions.

Elm name	Instances	Positions	Elm description	Pattern
CLV_NDR_NDR_1	RRG	34–36	N-Arg dibasic convertase (nardilysine) cleavage site (Xaa-|-Arg-Lys or Arg-|-Arg-Xaa)	.RK|RR[^KR]
CLV_PCSK_PC7_1	RTYGRRR	29–35	Proprotein convertase 7 (PC7, PCSK7) cleavage site (Arg-Xaa-Xaa-Xaa-[Arg/Lys]-Arg-|-Xaa)	[R]...[KR]R.
LIG_14-3-3_2	RGQSSLF	35–41	Longer mode 2 interacting phospho-motif for 14-3-3 proteins with key conservation RxxxS#p.	R..[^P][ST][IVLM].
LIG_BRCT_BRCA1_1	QSSLF	37–41	Phosphopeptide motif which directly interacts with the BRCT (carboxy-terminal) domain of the Breast Cancer Gene BRCA1 with low affinity	.S..F
LIG_Clathr_ClatBox_1	LFPME	40–44	Clathrin box motif found on cargo adaptor proteins, it interacts with the beta propeller structure located at the N-terminus of Clathrin heavy chain.	L[IVLMF].[IVLMF][DE]
LIG_PDZ_3	NDSI	6–9	Class III PDZ domains binding motif	.[DE].[IVL]
MOD_CK1_1	SDSSLLI	20–26	CK1 phosphorylation site	S..([ST])...
MOD_Cter_Amidation	YGRR	31–34	Peptide C-terminal amidation	(.)G[RK][RK]
MOD_GSK3_1	LNYSYPGS	13–20	GSK3 phosphorylation recognition site	...([ST])...[ST]
MOD_N-GLC_1	ENDSIK LNYSYP	5–10 13–18	Generic motif for N-glycosylation. Shakin-Eshleman et al. showed that Trp, Asp, and Glu are uncommon before the Ser/Thr position. Efficient glycosylation usually occurs when ~60 residues or more separate the glycosylation acceptor site from the C-terminus	.(N)[^P][ST]..
MOD_PKB_1	RRRGQSSLF	33–41	Pkb Phosphorylation site	R.R..([ST])...
MOD_PLK	SDSSLLI	20–26	Site phosphorylated by the Polo-like-kinase	.[DE].[ST][ILFWMVA]..
TRG_LysEnd_APsAcLL_1	DSSLLI	21–26	Sorting and internalisation signal found in the cytoplasmic juxta-membrane region of type I transmembrane proteins. Targets them from the Trans Golgi Network to the lysosomal-endosomal-melanosomal compartments. Interacts with adaptor protein (AP) complexes	[DER]...L[LVI]

Note: The numeric positions in column 3 are based on the following sequence investigated: IMRAENDSIK**GRLNYSYPGSDSSLLINARTYGRRR**G QSSLFPMED. The letters in bold represent the 25-aa (the 17-aa plus the 8-aa) region.

considered "fragmented" and removed from further analysis. Vector NTI (Align module, AlignX—Align selected molecules) and ClustalW2 (www.ebi.ac.uk/clustalw2/, at default settings) were used to align the alternatively spliced protein isoforms as well as human and mouse protein orthologs.

A phylogenetic tree was generated for Cpeb1–4 from multiple vertebrate species with the aid of Geneious Pro ver 4.8 (www.geneious.com) at the following settings: Tree Alignment Options—Cost matrix: Blosum62. Gap open penalty: 12. Gap extension penalty: 3. Alignment type: Global alignment with free end gaps. Tree Builder Options—Genetic distance model: Jukes-Cantor. Tree build method: Neighbor-joining. Outgroup: No Outgroup. The accession numbers of protein sequences used to generate the phylogenetic tree were listed in supplementary Table 10.

Functional prediction of alternatively used motifs

No 3D structural information is readily available except for the RRM structure for human CPEB3 (NCBI Cn3D database). Whereas the deletions may lead to critical changes in the secondary/tertiary structures of the proteins, we could only predict the possible functional motifs based on consideration of the linear sequences at this stage. Potential functional motifs were identified with the aid of Eukaryotic Linear Motif resource (ELM, http://elm.eu.org). Since the lengths of the functional motifs used by the ELM algorithm are within 10-aa, to keep the integrity of potential motifs, we included 10-aa N-terminal and 10-aa C-terminal to the regions encompassing the short deletions. Additional predicted and experimentally confirmed phosphorylation sites were from PhosphoSitePlus (www.phosphosite.org).

Abbreviations

Cpeb, mouse cytoplasmic polyadenylation element binding protein, cDNA, or used as a general term for the cDNA across species; Cpeb, mouse cytoplasmic polyadenylation element binding protein, protein, or used as a general term for the protein across species; *CPEB,* human cytoplasmic polyadenylation element binding protein, cDNA; CPEB, human cytoplasmic polyadenylation element binding protein, protein; CPE, cytoplasmic polyadenylation element; RRM, RNA recognition motif; CDS, protein coding sequence; UTR, untranslated region; 5' RACE, 5' rapid amplification of cDNA ends; NMR, nuclear magnetic resonance; MS, mass spectrometry; Mapk, mitogen-activated protein kinase; Pka, protein kinase a; Pkb, protein kinase b; Camk2a, calcium/calmodulin-dependent protein kinase 2 alpha; Hnrnp, heterogeneous nuclear ribonucleoprotein; Ptbp, polypyrimidine tract binding protein; Pabp2, poly (A) binding protein 2; ELM, eukaryotic linear motif.

Acknowledgments

We thank Dr. Ben Harrison and Dr. Eric Rouchka for helpful discussions. This study was supported by NEI R01EY017594, NCRR P20 RR16481 and NIEHS P30ES014443.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Hake LE, Richter JD. CPEB is a specificity factor that mediates cytoplasmic polyadenylation during Xenopus oocyte maturation. *Cell*. 1994;79:617–27.
2. Eliscovich C, Peset I, Vernos I, Mendez R. Spindle-localized CPE-mediated translation controls meiotic chromosome segregation. *Nat Cell Biol*. 2008;10: 858–65.
3. Groisman I, Huang YS, Mendez R, Cao Q, Theurkauf W, Richter JD. CPEB, maskin, and cyclin B1 mRNA at the mitotic apparatus: implications for local translational control of cell division. *Cell*. 2000;103:435–47.
4. Wu L, Wells D, Tay J, et al. CPEB-mediated cytoplasmic polyadenylation and the regulation of experience-dependent translation of alpha-CaMKII mRNA at synapses. *Neuron*. 1998;21:1129–39.
5. Theis M, Si K, Kandel ER. Two previously undescribed members of the mouse CPEB family of genes and their inducible expression in the principal cell layers of the hippocampus. *Proc Natl Acad Sci U S A*. 2003;100: 9602–7.
6. Wang XP, Cooper NG. Characterization of the transcripts and protein isoforms for cytoplasmic polyadenylation element binding protein-3 (CPEB3) in the mouse retina. *BMC Mol Biol*. 2009;10:109.
7. Kurihara Y, Tokuriki M, Myojin R, et al. CPEB2, a novel putative translational regulator in mouse haploid germ cells. *Biol Reprod*. 2003;69: 261–8.
8. Huang YS, Kan MC, Lin CL, Richter JD. CPEB3 and CPEB4 in neurons: analysis of RNA-binding specificity and translational control of AMPA receptor GluR2 mRNA. *EMBO J*. 2006;25:4865–76.
9. Vogler C, Spalek K, Aerni A, et al. CPEB3 is associated with human episodic memory. *Front Behav Neurosci*. 2009;3:4.
10. McGrew LL, Dworkin-Rastl E, Dworkin MB, Richter JD. Poly(A) elongation during Xenopus oocyte maturation is required for translational recruitment and is mediated by a short sequence element. *Genes Dev*. 1989; 3:803–15.

11. Paris J, Richter JD. Maturation-specific polyadenylation and translational control: diversity of cytoplasmic polyadenylation elements, influence of poly(A) tail size, and formation of stable polyadenylation complexes. *Mol Cell Biol*. 1990;10:5634–45.
12. Wilczynska A, Aigueperse C, Kress M, Dautry F, Weil D. The translational regulator CPEB1 provides a link between dcp1 bodies and stress granules. *J Cell Sci*. 2005;118(Pt 5):981–92.
13. Welk JF, Charlesworth A, Smith GD, MacNicol AM. Identification and characterization of the gene encoding human cytoplasmic polyadenylation element binding protein. *Gene*. 2001;263(1–2):113–20.
14. Ding J, Hayashi MK, Zhang Y, Manche L, Krainer AR, Xu RM. Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. *Genes Dev*. 1999;13:1102–15.
15. Kenan DJ, Query CC, Keene JD. RNA recognition: towards identifying determinants of specificity. *Trends Biochem Sci*. 1991;16:214–20.
16. Burd CG, Dreyfuss G. Conserved structures and diversity of functions of RNA-binding proteins. *Science*. 1994;265:615–21.
17. Crichlow GV, Zhou H, Hsiao HH, et al. Dimerization of FIR upon FUSE DNA binding suggests a mechanism of c-myc inhibition. *EMBO J*. 2008;27: 277–89.
18. Perez I, McAfee JG, Patton JG. Multiple RRMs contribute to RNA binding specificity and affinity for polypyrimidine tract binding protein. *Biochemistry*. 1997;36:11881–90.
19. Song J, McGivern JV, Nichols KW, Markley JL, Sheets MD. Structural basis for RNA recognition by a type II poly(A)-binding protein. *Proc Natl Acad Sci U S A*. 2008;105:15317–22.
20. Lorkovic ZJ, Barta A. Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant Arabidopsis thaliana. *Nucleic Acids Res*. 2002;30:623–35.
21. Dephoure N, Zhou C, Villen J, et al. A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci U S A*. 2008;105:10762–7.
22. Villen J, Beausoleil SA, Gerber SA, Gygi SP. Large-scale phosphorylation analysis of mouse liver. *Proc Natl Acad Sci U S A*. 2007;104:1488–93.
23. Zanivan S, Gnad F, Wickstrom SA, et al. Solid tumor proteome and phosphoproteome analysis by high resolution mass spectrometry. *J Proteome Res*. 2008;7:5314–26.
24. Munton RP, Tweedie-Cullen R, Livingstone-Zatchej M, et al. Qualitative and quantitative analyses of protein phosphorylation in naive and stimulated mouse synaptosomal preparations. *Mol Cell Proteomics*. 2007;6: 283–93.
25. Trinidad JC, Thalhammer A, Specht CG, et al. Quantitative analysis of synaptic phosphorylation and protein expression. *Mol Cell Proteomics*. 2008;7:684–96.
26. The UniGene Database [http://www.ncbi.nlm.nih.gov/unigene].
27. The UniProt Database [http://www.uniprot.org].
28. The UCSC BLAT [www.genome.ucsc.edu/cgi-bin/hgBlat].
29. The NCBI Database [http://www.ncbi.nlm.nih.gov/sites/entrez].
30. The Eukaryotic Linear Motif resource [http://elm.eu.org].
31. The PhosphoSitePlus [www.phosphosite.org].
32. ClustalW2 [www.ebi.ac.uk/clustalw2].
33. Geneious v4.8, Drummond AJ, Ashton B, et al. 2009. Available from http://www.geneious.com/
34. Current nomenclature guidelines by the HUGO Gene Nomenclature Committee at the European Bioinformatics Institute: http://www.genenames.org

Supplementary Materials

Table S1. Transcript and protein variants for mouse *Cpeb1*. Each mRNA corresponds to the protein in the same row. The translation is regarded as "disconnected" when none of the 6 frames gives continuous read. The translation is considered "fragmented" if it starts with the first codon which is not a Methionine, or terminates at the last codon which is not a stop codon. The asterisk * indicates that the protein sequence is not documented in any database but is derived from a theoretical translation of cDNA using Vector NTI software. cDNAs highlighted in grey are used for comparison.

mRNA	Size (bp)	Description	Protein	Size (aa)	Comments
AK077799.1	1737	Mus musculus adult male thymus cDNA	BAC37017.1	140 (422–561)	Fragmented
AK207851.1	402	Mus musculus cDNA		–	Disconnected translation
AK199617.1	379	Mus musculus cDNA		–	Disconnected translation
AK136088.1	2697	Mus musculus *in vitro* fertilized eggs cDNA product:cytoplasmic polyadenylation element binding protein 1, full insert sequence	Translated*	561	Possibly misreading of 1 nucleotide; otherwise gives identical product as NP_031781
AK135615.1	814	Mus musculus *in vitro* fertilized eggs cDNA, product:cytoplasmic polyadenylation element binding protein 1, full insert sequence		–	Disconnected translation
BC125476.1	1777	Mus musculus cytoplasmic polyadenylation element binding protein 1, mRNA, complete cds	AAI25477.1	562	1-aa insertion compared to NP_031781
BC144948.1	1759	Mus musculus cDNA	Translated*	556	5-aa deletion (PKGNM) compared to NP_031781
Y08260.1	2610	M.musculus mRNA for CPEB protein	CAA69588.1	561	
NM_007755.4	2612	Mus musculus cytoplasmic polyadenylation element binding protein 1 (*cpeb1*), mRNA	NP_031781.1	561	

Table S2. Across-paralog comparison of the exon patterns for *Cpeb1*. The numbers represent the lengths of exons in a sequential order. The locations and size of exons were determined by aligning the cDNA sequence to the genomic sequence. For simplicity, only one RefSeq sequence is used for each organism. The patterns in C. elegans and Drosophila are not consistent with those in the vertebrates. However the patterns among the vertebrates are highly conserved. The asterisk represents the variable region of ±15-nt discussed in the text, which would subsequently lead to 5-aa deletion or insertion.

C. elegans							415	360	111	91	340	216	757	NM_066650.4
Drosophila				419	500	220	482	137	232	719	181	386	1584	NM_079736.2
chicken		49	178	189	227	250	114	90	137	199	95	81	111	XM_413713.2
chimpanzee	62	1844	175	189	227	253	114	90	137	199	95	81	1504	XM_001158685.1
cow			189	189	227	253	114	90	137	199	95	81	1456	XM_864691.3
frog		91	178	201	218	253	114	105	137	199	95	81	1479	NM_001017330.2
horse			183	192	227	253	114	90	137	199	95	81	1463	XM_001498253.2
human		122	175	189	227	253	114	90	137	199	95	81	1502	NM_030594.3
marmoset		120	175	189	224	253	114	90	137	199	95	81	1469	XM_002749169.1
mouse		43	175	189	224	253	114	105	137	199	95	81	997	NM_007755.4
orangutan		256	175	189	227	253	114	90	137	199	95	81	1500	NM_001132960.1
pig			175	189	227	253	114	105	137	199	95	81	111	NM_001097510.1
rat		44	175	189	224	253	114	105	137	199	95	81	999	NM_001106276.1
zebrafish		30	169	192	209	253	114	105	137	199	95	81	1303	NM_131427.1
								*						

Table S3. Transcript and protein variants for mouse *Cpeb2*. Each mRNA corresponds to the protein in the same row. The translation is considered "fragmented" if it starts with the first codon which is not a Methionine, or terminates at the last codon which is not a stop codon. The asterisk * indicates that the protein sequence is not documented in any database but is derived from a theoretical translation of cDNA using Vector NTI software. cDNAs highlighted in grey are used for comparison.

mRNA	Size (bp)	Description	Protein	Size (aa)	Comments
AK076221.1	4536	Mus musculus 15 days embryo head cDNA,	Translated*	189	Fragmented
AK042065.1	2629	Mus musculus 3 days neonate thymus cDNA,	Translated*		30-aa deletion, 8-aa insertion compared to NP_787951
AB100307.1	1942	Mus musculus *cpeb2* mRNA, complete cds	BAC57076.1	521	
NM_175937.2	1942	Mus musculus cytoplasmic polyadenylation element binding protein 2 (c*peb2*), mRNA	NP_787951.1	521	
AK164866.1	904	Mus musculus 15 days embryo head cDNA,		89	Fragmented
AK154330.1	3576	Mus musculus NOD-derived CD11c +ve dendritic cells cDNA,		–	Disconnected translation
BC107349.1	1801	Mus musculus cytoplasmic polyadenylation element binding protein 2, mRNA	AAI07350.1	521	3-nt deletion in 5' UTR
BC107350.1	1801	Mus musculus cytoplasmic polyadenylation element binding protein 2, mRNA	AAI07351.1	521	3-nt deletion in 5' UTR

Table S4. Across-paralog comparison of the exon patterns for *Cpeb2*. The numbers represent the lengths of exons in a sequential order. The locations and size of exons were determined by aligning the cDNA sequence to the genomic sequence. For simplicity, only one RefSeq sequence is used for each organism. The pattern in C. elegans is not consistent with those in the vertebrates. However the patterns among the vertebrates are highly conserved. The asterisks represent the variable regions of ±90-nt and ±24-nt discussed in the text, which would subsequently lead to 30-aa and 8-aa deletion or insertion. The pound sign indicates a newly updated isoform in which the use of an alternative exon leads to an extra-long protein.

C. elegans								47	337	174	236	776	196	NM_062835.2
cow			295		91	51	24	171	90	119	115	182	3994	XM_001787297.1
horse			289		91	51	24	171	90	119	115	182	3992	XM_001498900.2
human (new)		1662	282		91	51		171	90	119	115	182	4001	NM_182646.2
human		438	282		91	51		171	90	119	115	182	3998	NM_182646.1
mouse (new)		1626#	282	90	91	51		171	90	119	115	182	4006	NM_175937.3
mouse	59	200	282	90	91	51		171	90	119	115	182	472	NM_175937.2
rat		203	265	90	91	51		171	90	140	115	182	4003	NM_001108361.1
zebrafish			224		91	51		174	90	119	115	182	307	NM_001177457.1
				*			*							

Table S5. Transcript and protein variants for mouse *Cpeb3*. Each mRNA corresponds to the protein in the same row. The translation is regarded as "disconnected" when none of the 6 frames gives a continuous read. The # indicates the discrepancy we found from a theoretical translation of cDNA using Vector NTI software. The asterisk * indicates that the protein sequence is not documented in any database but is derived from a theoretical translation of cDNA using Vector NTI software. cDNAs highlighted in grey are used for comparison.

mRNA	Size (bp)	Description	Protein	Size (aa)	Comments
AK044639.1	4223	Mus musculus adult retina cDNA	Translated*	693	*189th a top codon; 23-aa deletion
AK029261.1	2411	Mus musculus 0 day neonate head cDNA	Translated*	469	216-aa truncation; a 23-aa deletion; a 8-aa deletion
AB093274.1	5310	Mus musculus mRNA for mKIAA0940 protein	BAC41458.1#	716	#Reported as 722 aa, but we presume it is 716 aa identical to NP_938042, because the 1st Methionine is the 7th aa.
AY313774.1	3148	Mus musculus cytoplasmic polyadenylation element binding protein 3 (c*peb3*) mRNA, complete cds	AAQ20843.1	716	Compare to NP_938042: One amino acid conversion: 372nd N -> *P*
NM_198300.2	5792	Mus musculus cytoplasmic polyadenylation element binding protein 3 (c*peb3*), mRNA	NP_938042.2	716	
AK147243.1	5660	Mus musculus cDNA, cytoplasmic polyadenylation element binding protein 3, full insert sequence	BAE27791.1	716	
AK161513.1	2097	Mus musculus adult male testis cDNA, cytoplasmic polyadenylation element binding protein 3, full insert sequence	BAE36436.1	561	23-aa deletion; Early termination with conversion: VELA -> GEWK
BC128377.1	2142	Mus musculus cytoplasmic polyadenylation element binding protein 3, mRNA, complete cds	AAI28378.1	477	216-aa truncation; a 23-aa deletion;
BC128378.1	465	Mus musculus cDNA		-	Discontinued translation

Table S6. Across-paralog comparison of the exon patterns for *Cpeb3*. The numbers represent the lengths of exons in a sequential order. The locations and size of exons were determined by aligning the cDNA sequence to the genomic sequence. For simplicity, only one RefSeq sequence is used for each organism. The pattern in C. elegans is not consistent with those in the vertebrates. However, the patterns among the vertebrates are highly conserved. The asterisks * represent the variable regions of ±69-nt and ±24-nt discussed in the text, which would subsequently lead to 23-aa and 8-aa deletion or insertion. These two regions are alternatively spliced in several species. The # represents a predicted variable region of ±27-nt predicted based on the indication of variants among multiple organisms, and which may lead to a 9-aa insertion or deletion. The ## indicate that the first exon in human and chimpanzee (193-nt), when mapped to mouse genome, would locate just upstream of the first exon in mouse (61-nt). Computational translation indicates that if the 61-nt extends into and beyond the 193-nt, Cpeb3 protein would be continuously extended at the N-terminus. This extra-long isoform of Cpeb3, if proven real, would provide support for a Cpeb3 isoform which is larger than 100 kD as previously reported.

C. elegans						152	930	159	601	213	747	NM_059279.5
chimpanzee	193##	1019	160	57		141	90	119	115	182	1792	XM_001145135.1
cow		1012	160	57		141	90	119	115	182	3784	XM_875092.3
frog		1098	91	57	24	168	90	119	115	182	331	NM_001015925.2
horse	161	1010	160	57		141	90	119	115			XM_001917417.1
human	193##	1016	160	57		141	90	119	115	182	3800	NM_014912.4
mouse	61	1019	160	57	24	168	90	119	115	182	3797	NM_198300.2
rat		1209	160	57	24	168	90	119	115	182	345	XM_220043.5
zebrafish		814	85	57	24	168	90	119	115	182	252	NM_001167662.1
			*		*	#						

Table S7. Transcripts and protein variants for mouse *Cpeb4*. Each mRNA corresponds to the protein in the same row. The translation is regarded as "disconnected" when none of the 6 frames gives continuous read. The translation is considered "fragmented" if it starts with the first codon which is not a Methionine, or terminates at the last codon which is not a stop codon. The asterisk * indicates that protein sequence is not documented in the database but derived from translation of cDNA using Vector NTI software. cDNAs highlighted in grey are used for comparison.

mRNA	Size (bp)	Description	Protein	Size (aa)	Comments
AK089951.1	1774	Mus musculus kidney CCL-142 RAG cDNA	Translated*	128 (1–128)	Fragmented
AK088039.1	1904	Mus musculus 2 days neonate thymus thymic cells cDNA		–	Disconnected translation
AK079421.1	3217	Mus musculus adult male bone cDNA,		–	Disconnected translation
AY313775.1	2313	Mus musculus cytoplasmic polyadenylation element binding protein 4 (*cpeb4*) mRNA, complete cds	AAQ20844.1	729	
AK173229.1	7585	Mus musculus mRNA for mKIAA1673 protein	BAD32507.1	704	25-aa deletion
BC079599.1	5437	Mus musculus cytoplasmic polyadenylation element binding protein 4, mRNA	Translated*	339 (383–721)	8-aa deletion
AK162101.1	2164	Mus musculus *in vitro* fertilized eggs cDNA,	BAE36725.1	262 (1–262)	Fragmented
AK154289.1	2438	Mus musculus NOD-derived CD11c +ve dendritic cells cDNA	BAE32491.1	338 (1–338)	Fragmented
BC115431.1	2279	Mus musculus cytoplasmic polyadenylation element binding protein 4, mRNA	AAI15432.1	704	25-aa deletion
BC115430.1	2279	Mus musculus cytoplasmic polyadenylation element binding protein 4, mRNA	AAI15431.1	704	25-aa deletion
BC145865.1	2300	Mus musculus cytoplasmic polyadenylation element binding protein 4, mRNA	AAI45866.1	729	
BC145863.1	2300	Mus musculus cytoplasmic polyadenylation element binding protein 4, mRNA	AAI45864.1	729	
AK021394.1	1066	Mus musculus 0 day neonate eyeball cDNA	Translated*	141 (589–729)	Fragmented
AK015401.1	1586	Mus musculus adult male testis cDNA	BAB29832.1	295 (435–729)	
AK015381.1	1077	Mus musculus adult male testis cDNA	BAB29821.1	295 (435–729)	
NM_026252.3	2312	Mus musculus cytoplasmic polyadenylation element binding protein 4 (*cpeb4*), mRNA	NP_080528.2	729	

Table S8. Across-paralog comparison of the exon patterns for *Cpeb4*. The numbers represent the lengths of exons in a sequential order. The locations and size of exons were determined by aligning the cDNA sequence to the genomic sequence. For simplicity, only one RefSeq sequence is used for each organism. The patterns among the vertebrates are highly conserved. The asterisks * represent the variable regions of ±51-nt and ±24-nt discussed in the text, which would subsequently lead to 17-aa and 8-aa deletion or insertion. The question mark indicates that the size of the exon was unknown due to missing information in the genomic sequence. As a result, the alignment of the exons in grey to this map is uncertain.

chimpanzee	39			24	174	90	119	115	182	5091	XM_001155021.1
cow	203	82			174	90	119	115	182	4401	NM_001105420.1
dog	1682	82	51		174	90	119	115	182	4739	XM_536428.2
human	2531	82	51	24	174	90	119	115	182	4401	NM_030627.2
horse	1662	82	51	24	174	90	119	115	182	4421	XM_001502804.2
marmmoset	2541	82	51	24	174	90	119	115	182	4535	XM_002744562.1
mouse	1202	82	51	24	174	90	119	115	182	273	NM_026252.3
rabbit	1125	82	51	24	174	90	119	115	182	228	XM_002710388.1
rat	1125	82	51		174	90	119	115	182	892	NM_001106992.1
Rhesus	99	82			174	90	119	115	182	4231	XM_001097641.1
zebrafish				25	1172	93	384?	115	182	375	NM_200981.1
			*	*							

Table S9. Sequences and locations of primers used for *Cpeb1*, *Cpeb2*, and *Cpeb4* RT-PCR.

Gene	Primer name	Location	Sequence
Cpeb1	f 1_3	exon1/3	5′-GGCTTTCTCTCTGACTTCCAGGACTC-3′
	r 3	exon 3	5′-GACTGTGTGCTGCTCTGGGCTG-3′
	f 7	exon 7	5′-TGGTAAGGATGGCAAGCACCCC-3′
	r 8	exon 8	5′-ACGGACTTCTCTAGTTCAAACACCAAA-3′
Cpeb2	fwd	exon 3	5′-TCAGGACAGACAACAATAGTAACACA-3′
	rev	exon 8	5′-ATCTATTGGAAATAGGGAAGAGCGA-3′
Cpeb3	ex4 f	exon 4	5′-TGGATGGAGGATAACGCTTT-3′
	ex5 f	exon 5	5′-CTGACCATGAGCCTCTGAAA-3′
	ex8 d_r1	exon 8(–27nt)/6	5′-CATCCAAGAAGGCGTTGTTA-3′
	ex8 d_r2	exon 8(–27nt)/7	5′-TCCAAGAAGGCGTCTCGTC-3′
Cpeb4	f 2_3	exon 2/3	5′-AATGATTCCATTAAAGGTCGTCTA-3′
	f 2_4	exon 2/4	5′-AAATGATTCCATTAAAGCAAGG-3′
	f 2_5	exon 2/5	5′-AATGATTCCATTAAAGGTCAGTCT-3′
	r 5_3	exon 5/3	5′-GGAAACAATGAAGACTGACCATTA-3′
	r 5_4	exon 5/4	5′-GAAACAATGAAGACTGACCTCTCC-3′
	r 5	exon 5	5′-GAGGCAATCCACCCACAA-3′

Table S10. Protein sequences used to generate the phylogenetic tree. For simplicity, only the longest RefSeq sequences were used for each species. As in the tree, C1–C4 represents Cpeb1–Cpeb4.

C1	C2		C3		C4	
chicken	XP_413713.2	cow	XP_001787349.2	chimpanzee	XP_001145135.1	chimpanzee	XP_001155432.1
chimpanzee	XP_001158685.1	horse	XP_001498950.2	cow	XP_880185.2	human	NP_085130.2
cow	XP_869784.3	human	NP_872291.1	frog	NP_001015925.1	marmoset	XP_002744608.1
frog	NP_001017330.1	mouse	NP_787951.1	horse	XP_001917452.1	mouse	NP_080528.2
horse	XP_001498303.2	rat	NP_001101831.1	human	NP_055727.3	rat	NP_001100462.1
human	NP_085097.3	zebrafish	NP_001170928.1	mouse	NP_938042.2	zebrafish	NP_957275.1
marmoset	XP_002749216.1			rat	XP_220043.5		
mouse	NP_031781.1			zebrafish	NP_001161134.1		
orangutan	NP_001126432.1						
pig	NP_001090979.1						
rat	NP_001099746.1						
zebrafish	NP_571502.1						