Evolutionary history of Cytoplasmic Polyadenylation Element-Binding Proteins

Principal Investigator: Labib Rouhana and Joseph Ryan

Draft or Version Number: v.1.4

21 October 2022

LIST OF ABBREVIATIONS

<Insert text>	<Insert text>
CPEB	Cytoplasmic Polyadenylation Element-Binding Protein
RBP	RNA-Binding Protein
RRM	RNA Recognition Motif
RBD	RNA-Binding Domain (region of CPEB containing RRMs)
ZZ	Zinc (binding domain)
poly(A)	Polyadenosine

1 INTRODUCTION: BACKGROUND INFORMATION AND SCIENTIFIC RATIONALE

1.1 BACKGROUND INFORMATION

Cytoplasmic Polyadenylation Element-Binding Proteins (CPEBs) are pivotal regulators of gene expression during gametogenesis and the initial stages of embryonic development of bilaterians. These RNA-binding proteins recognize sequences (consensus UUUUA$_{1-2}$U) in the 3'untranslated region of specific mRNAs and recruit cytoplasmic poly(A)-polymerases that increase the poly(A)-tail length of substrate mRNAs and consequently their translation. CPEBs have not been studied in outside of bilaterians, but a recent study identified them as one of "25 groups of metazoan-specific genes that are essential across the Animal Kingdom" (Paps and Holland, 2018).

1.2 RATIONALE

CPEB orthologs corresponding to both CPEB1 and CPEB2 subgroups have been found in many bilaterian groups, but detailed phylogenetic analyses that have included non-bilaterian animals as well as closely related non-animal lineages have yet to be performed. The results of such analyses could shine light on the origin and evolution of cytoplasmic polyadenylation.

1.3 HYPOTHESES

Orthologs of CPEB1 and CPEB2 are present in the genomes of all animals and absent from the genomes of non-animals. The most identical proteins in non-animal genomes will be RNA-binding proteins with essential functions shared throughout Eukaryota. Metazoan proteins with sequence similarity with CPEBs, which are not CPEB orthologs or orthologs of proteins present in non-metazoan genomes, share biochemical and biological functions with CPEBs.

1.4 OBJECTIVES

1. To validate the presence or absence of CPEB orthologs in genomes of non-bilaterian animals.

2. To identify non-animal proteins that share a most-recent common genic ancestor with CPEBs.

3. To document the evolutionary history of CPEBs and related genes.

2 STUDY DESIGN AND ENDPOINTS

1) **Identify putative CPEBs within genomes:** Starting with the RBD (tandem RRMs and ZZ domain) of human CPEB1 (NP_001275748.1; AA 234-479) and CPEB4 (AAH36899.1; AA 66-309) protein sequences (Afroz et al., 2014), perform TBLASTN searches against:
Saccharomyces cerevisiae (Fungi),
Schizosaccharomyces pombe (Fungi),
Arabidopsis thaliana (Plant),
Capsaspora ocwazarki (Filasterea),
Salpingoeca rosetta (Choanoflagellatea),
Monosiga brevicollis (Choanoflagellatea),
Amphimedon queenslandica (Porifera),
Mnemiopsis leidyi (Ctenophora),
Nematostella vectensis (Cnidaria),
Trichoplax adhaerens (Placozoa),
Drosophila melanogaster (Arthropoda),
Capitella teleta (Annelida),
Schmidtea mediterranea (Platyhelminth),
Mus musculus (Vertebrata),
and *Danio rerio* (Vertebrata)
gene models to identify organisms with CPEB orthologs. The cut-off point is set to Evalue $\leq$ 0.5, because lower Evalues (10e-6) would not allow for inclusion of non-orthologous homologs of CPEBs. An additional step to remove redundant sequences is performed by keeping only the longest isoform.

2) **Align domains in CPEB homologs:** We will search the putative CPEBs and outgroups identified above with the following HMMs from PFAM: RRM_1 (PF00076), RRM_7 (PF16367), and CEBP_ZZ (PF16366). For CEBP_ZZ, which should only occur once per protein, we will create an alignment to the CEBP_ZZ HMM using the hmm2aln.pl script which runs hmmsearch, stockhom2fasta, and custom code to remove indels. For the RRM_1 and RRM_7 HMMs, which will often match the same domains, we will run hmmsearch separately with each of these HMMs and keep all non-overlapping results, and merge overlapping results by taking the lowest N-terminal coordinate and the highest C-terminal coordinate from the 2 results. We will extract the amino acid sequences between the start and end of the match and use mafft with default parameters to generate an alignment of RNA recognition motifs (RRMs).

3) Infer phylogenetic relationships among CPEBs and non-CPEB sequences (with similarity to CPEBs) by estimating gene trees of RBD sequence databases from step (1) and step (2) using IQTREE.

4) To understand the evolution of CPEB genes within ctenophores run a focused phylogenetic analysis on CPEBs within Ctenophora. Identify reciprocal best BLAST hits of *Mnemiopsis* CPEBs in the protein models of two other ctenophores (*Hormiphora californensis* and *Beroe ovata).* Align these with MAFFT (default parameters) and run a maximum-likelihood analysis using IQ-TREE (default parameters with automatic model finding and 1000 bootstraps).

3 LITERATURE REFERENCES

Paps, J & Holland, PW, 2018, 'Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty'. *Nature Communications*, vol 9.
Afroz T, Skrisovska L, Belloc E, Guillén-Boixet J, Méndez R, Allain FH. A fly trap mechanism provides sequence-specific RNA recognition by CPEB proteins. *Genes Dev*. 2014;28(13):1498-1514. doi:10.1101/gad.241133.114

4 PHYLOTOCOL AMENDMENT HISTORY

Version	Date	Significant Revisions
1.1	Dec. 19, 2020	Inclusion of E-value cut-off ≤ 0.1 for step 1
1.1	Dec. 19, 2020	Removal of steps 2, 3, 5 and 6 from previous version (1.0) because non-orthologous sequences will be identified in step 1 under the updated cut-off E=value.
1.2	Dec. 19, 2020	Previous to running tree phylogeny, HHM search showed that C. elegans sequences were highly derived (e.g. FOG-1 CEBP_ZZ domain was split between two predictions because the domain was 3x the size of the model.
1.2	Dec. 19, 2020	We know that there are at least two CPEBs (Rouhana et al. 2018) so we needed a different source. We BLASTed several gene predictions from SmedGD and PlanMine and the PlanMine hits were the most reasonable. However, We could not download the entire set of data, so we just downloaded the BLAST hits. The predicted protein products or corresponding NCBI protein sequences of these hits are in this file: Planmine_CPEB_homologs.fa
1.2	Dec. 21, 2020	We realized that E-value cut-off of < 0.5 (and not 0.1) was used for step 1. Corresponding correction added to the phylotocol.
1.3	Sept. 27, 2022	Corrected typo on point number 3 of study design, which referred to steps in a previous version of the phylotocol that were absent in version 1.2.
1.3	Sept. 27, 2022	Removal of step 4 of Study Design, as well as references to RAxML and MrBayes in step 3.
1.3	Sept. 27, 2022	Removed lines under "Work Completed So Far W Dates" which were only template.
1.4	Oct. 21, 2022	Added #4 to study design to run phylogenetic analysis on ctenophores only.