Computational Prediction of Genes Translationally Regulated by Cytoplasmic Polyadenylation Elements

Eric C. Rouchka[1], Xiangping Wang[2], James H. Graham[3], and Nigel G.F. Cooper[2]

[1] Department of Computer Engineering and Computer Science, J.B. Speed School of Engineering, 123 JB Speed Building, University of Louisville, Louisville, Kentucky USA
eric.rouchka@louisville.edu
[2] Department of Anatomical Sciences and Neurobiology, 500 South Preston Street, University of Louisville, Louisville, Kentucky USA
x0wang04@louisville.edu, nigelcooper@louisville.edu
[3] Department of Electrical and Computer Engineering, J.B. Speed School of Engineering, University of Louisville, Louisville, Kentucky USA
jhgrah01@louisville.edu

Abstract. Cytoplasmic post-transcriptional modification of mRNA transcripts in the form of polyadenylated (poly(A)) tails plays a key role in their translational control. The timing and degree of polyadenylation has been shown to be due in part to a consensus nucleotide sequence -- cytoplasmic polyadenylation elements (CPEs) which can be detected by a polyadenylation element binding protein (CPEB). An individual mRNA transcript controlled by CPEB may contain one or more CPE sites occurring upstream of a consensus hexamer poly-(A) signal. A probabilistic model, CPEDetector, is presented for predicting whether or not a gene's translation is mediated by CPEB. CPEDetector takes into account detected CPE sites, poly-A sites, and distance metrics between the detected locations. This approach is tested against the 3' untranslated regions (UTRs) of known genes using the UTRdb database.

Keywords: CPE, CPEB, bioinformatics, hidden Markov model, context free grammar, untranslated region.

1 Introduction

1.1 Central Dogma and Gene Regulation

The Central Dogma of Molecular Biology in summary states that the process of creating a protein encoded by a gene first begins with the genomic DNA, which is transcribed into an RNA intermediary template, known as messenger RNA (mRNA), which in turn is translated into a protein sequence using the genetic code. Regulation of genes can occur at either the transcriptional level, through complexes that form at transcription factor binding sites, or at the translational level. Transcription factor binding sites are typically found upstream (5') of the transcription start site (TSS). In contrast, many translational control regulators bind to the 3' untranslated region (UTR) downstream of the coding region. One such translational control mechanism is cytoplasmic polyadenylation.

S. Rajasekaran (Ed.): BICoB 2009, LNBI 5462, pp. 353–361, 2009.

1.2 Cytoplasmic Polyadenylation and Translational Control

In a number of molecular processes such as oogenesis, embryogenesis, and synaptic plasticity of the central nervous system, it is important to have the mRNA available temporally and spatially for specific and efficient protein production. One important mechanism underlying such translational control is cytoplasmic polyadenylation.

It has been long known that all eukaryotic pre-mRNAs undergo polyadenylation in the nucleus before they are exported to the cytoplasm. The newly synthesized pre-mRNA is endonucleolytically cleaved at about 10 nucleotides upstream of the polyadenylation signal (PAS), a hexamer sequence found in the 3' UTR. The poly(A) stretch (typically around 30nt in length) is then added to the newly formed 3' end. The poly(A) tail and its bound proteins are important for termination of transcription, export of the mRNA from the nucleus, and protection of the mRNA from degradation by exonuclease. Multiple variants of the PAS sequence have been identified in nature with different occurrences and distinct efficiencies for polyadenylation (table 1) [1].

Table 1. Alternative hexamer polyadenylation site (PAS) patterns

Hexamer	Frequency	Hexamer	Frequency
AAUAAA	0.6431	UUUAAA	0.0133
AUUAAA	0.1645	AAGAAA	0.0122
UAUAAA	0.0354	AAAAAG	0.0088
AGUAAA	0.0298	AAUGAA	0.0088
AAUAUA	0.0188	AAUAGA	0.0077
GAUAAA	0.0144	ACUAAA	0.0066
CAUAAA	0.0144	AAAACA	0.0055
AAUACA	0.0133	GGGGCU	0.0033

The short poly(A) tails added during nuclear polyadenylation helps to stabilize the mRNA and prevent it from degradation. However, the elongation of poly(A) tails in the cytoplasm (cytoplasmic polyadenylation) has a rather different role: to recruit the mRNA for translation. Cytoplasmic polyadenylation was first identified in eggs and single-cell embryos [2–5], where little RNA transcription activity was detected. Polyadenylation was immediately followed by translation. Timed expression of pre-stored mRNA has also been observed during early embryonic development to initiate mitosis and in some circumstances, to dictate the polarity of the embryo. Such translational activation is accompanied by elongation of the poly(A) tails [6,7]. The

Table 2. Alternative cytoplasmic polyadenylation element site (CPE) patterns

CPE Pattern
UUUUUAU
UUUUUGU
UUUUUACU
UUUUUGUU
UUUUAAU
UUUUACU
UUUUAUU