Small Open Reading Frames

Tiny Treasures of the Non-coding Genomic Regions

A Yazhini

Open Reading Frames (ORFs) are the DNA sequences in the genome that has the potential to be translated. Generally, only long ORFs (≥ 300 nucleotides or nt) are thought to be protein coding regions and are considered as genes in the genome annotation pipeline. Until recent years, small ORFs (smORFs) of less than 100 codons (< 300 nt) were regarded as non-functional on the basis of empirical observations. However, recent work on ribosome profiling and mass spectrometry have led to the discovery of many translating functional small ORFs and presence of their stable peptide products. Further, examples of biologically active peptides with vital regulatory functions underline the importance of smORFs in cell functions. Genome-wide analysis shows that smORFs are conserved across diverse species, and the functional characterization of their peptides reveals their critical role in a broad spectrum of regulatory mechanisms. Further analysis of small ORFs is likely throw light on many exciting, unexplored regulatory mechanisms in different developmental stages and tissue types.

A Yazhini is a research student at Molecular Biophysics Unit, Indian Institute of Science. She works mainly on protein evolution, protein structure prediction and structure of macromolecule assemblies under the guidance of Prof. N Srinivasan. Overall, her research revolves around structural and mechanistic understanding of proteins.

1. Introduction

Gene is the protein encoding functional unit in the genome. It consists of promoter, protein coding (exon) and terminal regions. The protein coding regions of the gene are called the 'open reading frames' (ORFs). It contains series of 3 nucleotide (nt) codons which determines the sequence of amino acids (aa) in a protein. Depending on the codon sequence and the translation initiation position in the ORFs, the amino acid composition is determined. The advent of next-generation sequencing technology has led to

Keywords
Gene evolution, non-coding RNA, small open reading frames, smORF–encoded peptides, next-generation sequencing, RNA-seq, ribosome profiling.

Recently, ORFs with less than 100 codons, called small ORFs have been identified from what was originally believed as the non-coding regions of the genome. smORFs are functional and are seen to produce biologically significant peptides.

enormous growth in whole genome sequencing. Hence today, genome sequence of many organisms are available for functional characterization. In genome annotation, one of the functional characterization methods is gene identification by looking at the presence of ORFs. In general, ORF length of at least 100 codons (300 nt) – an arbitrary cut-off – is considered for gene identification, and ORFs of less than 100 codons are excluded. Also, it is observed that many long non-coding regions in the genome contain such short length ORFs, and they are usually ignored during functional analysis as such short stretches are deemed as random occurrences [1]. Recently, ORFs with less than 100 codons, called small ORFs (smORFs) have been identified from what was originally believed as the non-coding regions of the genome. smORFs are functional and are seen to produce biologically significant peptides [2, 3]. Few functional analyses have elucidated that smORFs are involved in gene expression regulation, and peptides from them showed essential role in many biological pathways [4].

This article deals with the presence of smORFs in what is believed to be the non-coding regions of genome, difficulties in distinguishing functional smORFs from non-functional randomly occurring short stretches of DNA, and its biological significance in a wide range of cellular functions.

2. smORFs in lncRNAs

Long non-coding RNAs (lncRNAs) are RNA transcripts from non-coding regions which are only functional as an RNA and do not undergo translation to produce proteins. Unlike small RNAs such as miRNAs, siRNAs and snRNAs, lncRNAs are longer with at least 200 nt. They consist of 5'UTR[1], ORFs, 3'UTR and introns (in the case of eukaryotes) [1, 5]. Generally, ORFs in lncRNAs are very small, less than 100 codons (300 nt). It was assumed that these smORFs do not undergo translation and such short stretches could easily occur by chance. Recently, functional characterization of lncRNAs evidenced that peptides are synthesized from smORFs located within lncRNAs and showed biologically sig-

[1] UTR refers to untranslated region.

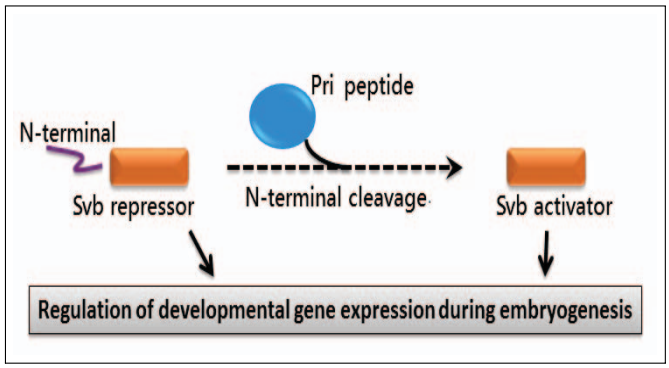

Figure 1. Regulatory function of Pri peptide during fruit fly embryogenesis.

nificant regulatory role . For example, in fruit flies, polished rice (Pri) or tarsal-less (Tal) – a 11 aa short peptide produced from lncRNA – promotes the activation of Svb (Shavenbaby) transcription factor by triggering the N-terminal cleavage of Svb repressor (*Figure* 1). This activation of Svb is essential for controlled gene expression during embryogenesis [6].

Apart from lncRNA mediated gene regulation, translation of smORFs from lncRNAs shows vital functions in cell survival under various stress conditions.

Similarly, during the embryo development of zebrafish, 58 aa long toddler polypeptide produced by a short ORF within a lncRNA, regulates cell migration during the gastrulation period [7]. Apart from lncRNA mediated gene regulation, translation of smORFs from lncRNAs have also showed vital functions in cell survival under various stress conditions. It was observed that smORFs facilitate localization of lncRNAs in the cytoplasm and enhance their translation [8].

3. smORFs in mRNAs

Messenger RNAs (mRNAs) undergo translation and produce functional proteins. They consist of ORFs (canonical) with minimum 100 codons flanked by 5′UTR and 3′UTR. Interestingly, smORFs are observed to be present in the UTR regions of mRNAs. Based on their location, smORFs are classified as uORFs, dORFs and sORFs (*Figure* 2).

Functional characterization of smORFs in the UTR regions has revealed that they are responsible for the regulation of translation

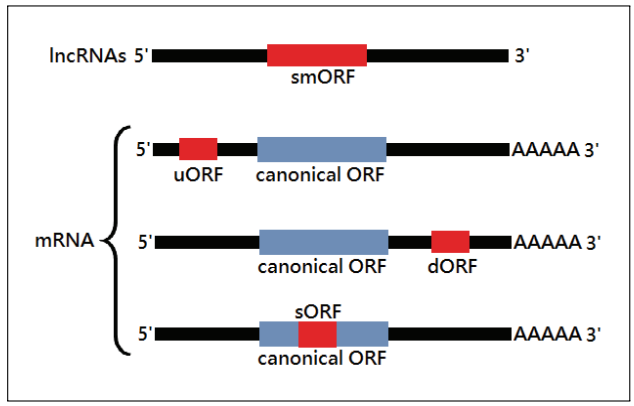

Figure 2. Classification of smORFs based on their location in the non-coding regions of the genome. uORF–smORF in 5′UTR; dORF–smORF in 3′UTR; sORF–smORF that overlaps with canonical long ORF in different reading frame.

of canonical ORFs [9]. In *Arabidopsis*, regulation of translation of canonical ORFs by 5′UTR smORFs is elucidated as follows (also see *Figures* 3, 4 and 5):

a) Regulated canonical ORF translation and ribosome stalling

Translation of uORFs negatively regulates the translation rate of canonical ORFs [10]. smORF-encoded peptides cause ribosomal stalling which bring down the rate of translation of canonical ORFs [11].

b) Regulated reinitiation

Environmental factors or metabolic changes cause translation reinitiation of canonical ORFs following the translation of uORFs [12].

c) Regulated mRNA decay

Transcription of uORF followed by canonical ORF leads to the synthesis of long mRNA containing uORF and canonical ORF

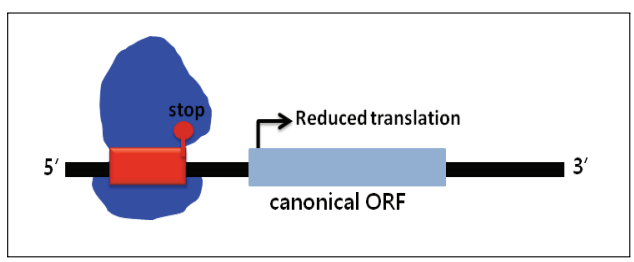

Figure 3. uORF mediated canonical ORF translation. Translation of canonical ORF is reduced by actively translating uORF located in the upstream region.

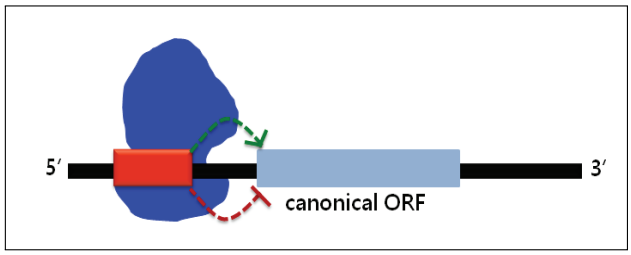

Figure 4. Reinitiation of translation at canonical ORF start codon. uORF controls the initiation of canonical ORF translation in response to the presence and absence of environmental factors.

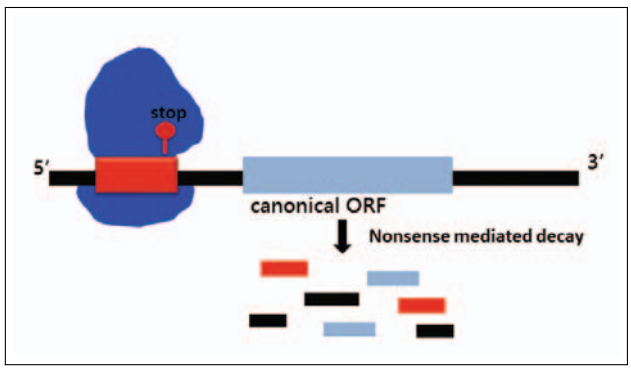

Figure 5. uORF mediated mRNA decay. Stability of mRNA is controlled by the uORF of canonical ORF.

regions. Presence of uORF region introduces the stop codon at 5′ end of the mRNA and mark it as an aberrant transcript. Such mRNA with premature stop codon is recognized by nonsense mediated decay (NMD) and gets degraded immediately. By this mechanism, uORF regulates mRNA viability [13].

4. Difficulties in Identifying smORFs

Genetic screening and mutagenesis[2] are the general methods used to identify and characterize functional regions of the genome. However, introduction of random mutations to smORFs is a very rare event, and hence it is not possible to detect smORFs through mutation studies. Several findings have showed that peptides from smORFs are expressed at specific developmental stages and are restricted to specific tissue types [14, 15]. However, their low abundance and small size make them extremely difficult to detect by mass spectrometry because they may get washed out during sample preparation. Conservation analysis by homologue search

[2] Random introduction of mutations.

methods like BLAST is also not suitable for smORFs since these methods are biased towards sequence length. If a query is very short (less than 100 aa), there is a high chance of detecting false positives with significant e-values. Analysis of randomly computer generated nucleotide sequences showed that occurrence of ORFs with less than 100 codons could be by chance and most of the protein coding regions have at least 100 codons in their ORFs [1, 15]. Thus, all the gene prediction methods (e.g., Glimmer, GeneMark and GenSCAN) are restricted to define transcript as mRNA only if ORF length is $\geq$ 100 codons. Also, 100 codon cut-off is used as the main criterion to distinguish mRNA from lncRNA in transcript expression analysis. So far, researchers thought that smORFs in lncRNA do not produce proteins and have been ignoring them in genome annotation pipeline. The recent discovery of high-throughput transcript expression techniques have evidenced that lncRNAs indeed undergo translation and produce stable peptides from their smORFs [8, 16, 17, 18].

The recent discovery of high-throughput transcript expression techniques have evidenced that lncRNAs indeed undergo translation and produce stable peptides from their smORFs.

5. Techniques for Identification of smORFs

RNA-sequencing (RNA-seq) is a whole transcriptome sequencing technique used to quantitatively measure the entire RNA transcript expression under the given cell conditions. In this experiment, fragments per kilobase of transcript per million mapped reads (FPKM) is used to quantify RNA transcript expression and is applicable for identifying short RNA transcripts from smORFs [16]. In 2012, ribosome profiling (Ribo-seq) technique was discovered to holistically identify translating RNA transcripts. It captures the snapshot of all the translating RNA regions [2, 16, 17]. Hence, Ribo-seq identifies truly translating smORFs which indirectly indicate their protein synthesizing ability. Further, functional importance of smORFs is examined by their conservation across species. There are many evolutionary conservation scores developed to find conserved elements in the genome. For example, PhastCons (PHylogenetic Analysis with Space/Time models Conservation score) is a phylo-HMM based method to analyze nucleotide sequence conservation using multiple genome

sequence alignment. High PhastCon score indicates that the genomic region is highly conserved among species. dN/dS or Ka/Ks[3] analysis is another conservation scoring method to analyze conservation at protein sequence level. It is calculated by taking the ratio of non-synonymous mutation rate to synonymous mutation rate in the given genomic region. Lower the score, higher is the protein sequence conservation with preference for synonymous mutations over non-synonymous mutations [2, 8, 18]. Although the above-mentioned scoring methods are applicable for the conservation analysis of smORFs, another method named PhyloCSF (Phylogenetic Codon Substitution Frequencies) was developed to accurately analyze conservation of smORFs as well as their protein coding potential. It considers the codon frequency, composition of genomic region, transition (A↔T or G↔C) and transversion (A/T↔G/C) rate and dN/dS ratio to calculate phyloCSF score. It examines whether a given genomic region is conserved as well as has coding potential or not. High phyloCSF score indicates that the region is well conserved and has potential to encode protein [19]. All aforementioned experiments and conservation scoring methods are currently applied to identify functional smORFs in whole genome analysis.

[3] dN/dS (non-synonymous to synonymous rate ratio) or Ka/Ks is the ratio of substitution rates to quantify evolutionary pressure on protein coding region.

6. Applications

Genome-wide smORFs identification has been carried out in model organisms such as *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (roundworm), *Danio rerio* (zebrafish), *Mus musculus* (mouse) and *Homo sapiens* (human). Analysis of such predicted smORFs indicates that they are conserved across species and prefer synonymous mutations over non-synonymous mutations to preserve the protein sequence composition [18]. Peptides from smORFs are involved in protein-protein interactions and regulate signaling and metabolic pathways [19]. For example, SgrT, a 43 aa long peptide binds to glucose transporter membrane protein and negatively regulates glucose influx into cytoplasm in *E. coli* under glucose toxicity condition [20]. Similarly, in mouse, myoregulin (MLN), phospholamban (PLN) and sar-

Analysis of smORFs indicates that they are conserved across species and prefer synonymous mutations over non-synonymous mutations.

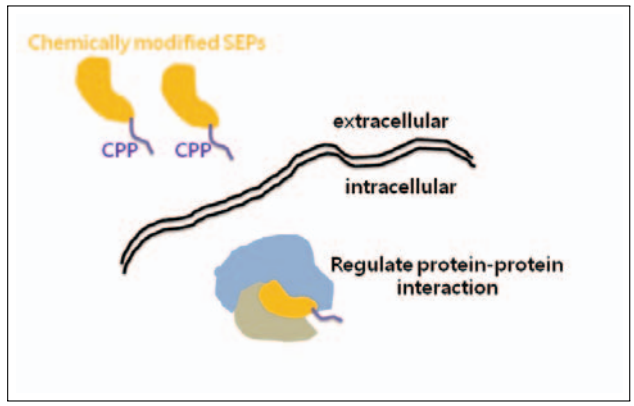

Figure 6. SEPs in therapeutics. Chemically modified SEPs can potentially be used as drugs for modulating protein-protein interactions as well as drug delivery system.

colipin (SLN) peptides interacts with the sarcoplasmic reticulum Ca^{2+} ATPase (SERCA) ion channel to regulate muscle contraction and relaxation [21]. Increasing functional characterization of smORF-encoded peptides (SEPs) in recent times reveals that they play a vital role in a broad range of biological regulatory mechanisms [3, 4, 6, 7, 20]. Structural characterization of predicted SEPs have shown that they are mostly disordered peptides and have protein interacting motifs. In humans, a 24 aa long peptide called humanin synthesized from smORF present in the non-coding RNA is involved in apoptosis by interacting with BAX (Bcl-2-associated X protein) [22]. Since SEPs are involved in protein-protein interactions and essentially regulate disease related pathways, they can be considered as potential drug targets. Nonetheless, a major problem in using peptides as a drug is that they have poor stability and they can be readily cleaved by proteases. Chemically modified stable SEPs with improved pharmacokinetic properties can be used as a drug to modulate protein-protein interactions (*Figure* 6) and also cell-penetrating peptide (CPP) tagged SEPs can act as potential drug delivery systems [23].

Potential applications of smORF-encoded peptides include the development of new therapeutic approaches to treat diseases efficiently.

Hence, potential applications of smORF-encoded peptides include the development of new therapeutic approaches to treat diseases efficiently.

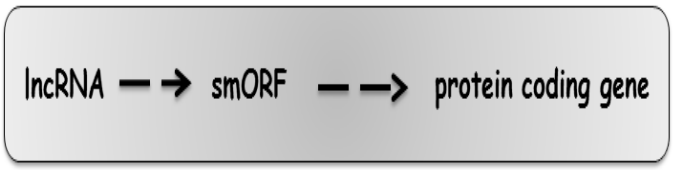

Figure 7. Gene evolution from non-coding regions in the genome.

7. Conclusion

High throughput identification of smORFs in well-annotated genomes has evidenced that functional smORFs are indeed present in lncRNAs and putative non-coding regions of mRNAs. RNA-seq and Ribo-seq techniques also show that smORFs undergo transcription followed by translation to produce stable peptides. Further functional characterization has to be carried out experimentally for individual cases to understand their biological importance. Their role in the regulation of gene expression, signaling pathways and tissue/developmental stage specific expression indicates that smORFs are highly essential for cell functions and could be used to develop new therapeutics. In addition, the presence of smORFs in the non-coding regions provides clue on the theory behind gene evolution. It is believed that gene evolution occurs in two ways: (1) editing already existing gene to make new protein coding genes and (2) protein/peptide coding gene emerging from non-coding region (*Figure* 7). smORFs in lncRNAs indicates that lncRNAs may be a birth pool of protein coding genes by initially acquiring smORFs and extending them during the course of evolution to make new protein coding genes [24].

Further analysis of such interesting genomic features will improve our understanding of cellular functions and gene evolution.

Acknowledgement

The author would like to thank Prof. N Srinivasan for his valuable suggestions and continuous encouragement.

Suggested Reading

[1] ME Dinger, KC Pang, TR Mercer and JC Mattick, Differentiating Protein-coding and Non-coding RNA: Challenges and Ambiguities, *PLoS Comput Biol*., Vol.4, No.11, p.e1000176, 2008.

[2] AA Bazzini *et al*., Identification of Small ORFs in Vertebrates Using Ribosome Footprinting and Evolutionary Conservation, *EMBO J*., Vol.33, No.9, pp.981–93, 2014.

[3] MM Kessler *et al*., Systematic Discovery of New Genes in the *Saccharomyces cerevisiae* Genome, *Genome Res*., Vol.13, No.2, pp.264–71, 2003.

[4] JP Albuquerque *et al*., Small ORFs: A New Class of Essential Genes for Development, *Genet Mol Biol*., Vol.38, No.3, pp.278–28, 2015.

[5] JS Mattick and IV Makunin, Non-coding RNA, *Hum Mol Genet*., Vol.15, (suppl_1): R17–R29, 2006.

[6] MI Galindo *et al*., Peptides Encoded by Short ORFs Control Development and Define a New Eukaryotic Gene Family, *PLoS Biol*., Vol.5, No.5, p.e106, 2007.

[7] A Pauli *et al*., Toddler: An Embryonic Signal That Promoted Cell Movement via Apelin Receptors, *Science*, Vol.343, p.1248636, 2014.

[8] Z Ji Z, R Song, A Regev and K Struhl, Many lncRNAs, 5′UTRs, and Pseudogenes are Translated, and Some are Likely to Express Functional Proteins, *eLife*, Vol.4, p.e08890, 2015.

[9] RP Hellens *et al*., The Emerging World of Small ORFs, *CellPress Trends in Plant Science*, Vol.21, No.4, pp.317–328, 2016.

[10] AG von Arnim, Q Jia, and JN Vaughn, Regulation of Plant Translation by Upstream Open Reading Frames, *Plant Sci*., Vol.214, pp.1–12, 2014.

[11] K Ito K and S Chiba S, Biological Significance of Nascent Polypeptides that Stall the Ribosome, *Regulatory Nascent Polypeptides*, Springer, pp.3-20, 2014.

[12] AE Firth and I Brierley I, Non-canonical Translation in RNA Viruses, *J Gen Virol*., Vol.93, No.7, pp.1385–409, 2012.

[13] O Shaul, Unique Aspects of Plant Nonsense-mediated mRNA Decay, *Trends Plant Sci*., Vol.20, No.11, pp.767–79, 2015.

[14] Y Kageyama, T Kondo and Y Hashimoto, Coding vs. Non-coding: Translatability of Short ORFs Found in putative Non-coding Transcripts, *Biochimie*, Vol.93, No.11, pp.1981–1986, 2011.

[15] S J Andrews and J A Rothnagel, Emerging Evidence for Functional Peptides Encoded by Short Open Reading Frame, *Nat Rev Genet*., Vol.15, No.3, pp.193–204, 2014.

[16] G L Chew *et al*., Ribosome Profiling Reveals Resemblance Between Long Non-coding RNAs and 5′ Leaders of Coding RNAs, *Development*, Vol.140, No.13, pp.2828–34, 2013.

[17] NT Ingolia *et al*., The Ribosomal Profiling Strategy for Monitoring Translation in Vivo by Deep Sequencing of Ribosome-Protected mRNA Fragments, *Nat Protoc*., Vol.8, pp.1534–50, 2012.

[18] SD Mackowiak *et al*., Extensive Identification and Analysis of Conserved Small ORFs in Animals, *Genome Biology*, Vol.16, No.179, 2015.

[19] MF Lin , I Jungreis and M Kellis, PhyloCSF: a Comparative Genomics Method

to Distinguish Protein Coding and Non-coding Regions, *Bioinformatics*, Vol.27, No.13, pp.i275–i282, 2011.

[20] C S Wadler and C K Vanderpool, A Dual Function for a Bacterial Small RNA: SgrS Performs Base Pairing-Dependent Regulation and Encodes a Functional Polypeptide, *Proc Natl Acad Sci*., USA., Vol.104, No.51, pp.20454–9, 2007.

[21] D M Anderson *et al*., A Micropeptide Encoded by a Putative Long Non-coding RNA Regulates Muscle Performance, *Cell*, Vol.160, No.4, pp.595–606, 2015.

[22] B Guo *et al*., Humanin Peptide Suppresses Apoptosis by Interfering with Bax Activation, *Nature*, Vol.423,No.6938, pp.456–61, 2003.

[23] A Saghatelian and J P Couso, Discovery and Characterization of smORF-encoded Bioactive Polypeptides, *Nat Chem Biol*., Vol.11, No.12, pp.909–16, 2015.

[24] C Xie *et al*., Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding RNAs, *PLoS Genet*., Vol.8, No.9, p.e1002942, 2012.

Address for Correspondence
A Yazhini
Molecular Biophysics Unit
Indian Institute of Science
Bangalore 560 012, India.
E-mail: yazhini@iisc.ac.in