Novel initiation codons and their role in protein isoform generation

AUGmenting the proteome

Mark J. Coldwell and **Joanne L. Cowan** (University of Southampton, UK)

As the field of molecular biology developed, and the understanding of how inherited genetic material results in the expression of proteins was established, the initial hypothesis was that one gene gave rise to one protein[1]. As researchers delved deeper into the organization of the genetic code and advances in messenger RNA (mRNA) and protein sequencing were subsequently made, it has become abundantly clear that multiple mechanisms exist meaning that many mRNAs encode more than one version of a protein. Although alternative promoters and alternative splicing play a considerable role in the generation of protein isoforms, in this article we discuss how usage of alternative translation initiation codons in eukaryotes can also lead to an expanded proteome.

Controlling initiation of protein synthesis

Decoding of eukaryotic mRNA by ribosomes during translation is divided into three stages: initiation, elongation and termination. In the first of these, the ribosome is positioned at the initiation codon of the mRNA in a process requiring a number of accessory factors (including the eukaryotic initiation factors or eIFs). These are subject to numerous controls and therefore initiation is considered to be a key control point of translation. There are numerous in-depth reviews on the subject, and these controls (such as sequestration of the mRNA cap-binding protein, eIF4E or the assembly of ribosomal subunits) are generally considered to govern the global rate of translation[2–4]. Another heavily regulated part of the translation initiation pathway is the recycling of eIF2. This acts as part of a ternary complex with GTP and initiating methionyl-tRNA (Met-tRNA$_i$). eIF2 discriminates between the initiator and elongator forms of Met-tRNA and in turn binds to the 40S ribosome, thus supplying the first tRNA and therefore amino acid to the nascent polypeptide chain. During initiation, the GTP molecule within the ternary complex is hydrolysed to GDP, and eIF2 and other eIFs are released to allow elongation of the protein to proceed. eIF2·GDP then undergoes recycling by the guanine-nucleotide-exchange factor (GEF) eIF2B to eIF2·GTP, which binds to Met-tRNA$_i$ ready for another initiation round. The recycling process can be inhibited by phosphorylation of eIF2 on the α subunit by four kinases[5], which are activated in response to cellular stresses, e.g. amino acid availability [general control non-derepressible 2 (GCN2)], presence of double-stranded RNA [dsRNA-dependent protein kinase (PKR)], an excess of haem compared with globin in red blood cell precursors [haem regulated inhibitor kinase (HRI)], or unfolded proteins in the endoplasmic reticulum [PKR-like endoplasmic reticulum kinase (PERK)]. The phosphorylation of eIF2 increases its binding to eIF2B, which leads to recycling being blocked, hence

Abbreviations: aORF, annotated ORF; eIF, eukaryotic initiation factor; eORF, extended ORF; GCN2, general control non-derepressible; HRI, haem regulated inhibitor kinase; IRES, internal ribosome entry site; Met-tRNA$_i$, initiating methionyl-tRNA; mRNA, messenger RNA; NLS, nuclear localization signal; ORF, open reading frame; PERK, PKR-like endoplasmic reticulum kinase; PKR, dsRNA-dependent protein kinase; rRNA, ribosomal RNA; tORF, truncated ORF; uORF, upstream ORF; UTR, untranslated region.

Key words: *alternative translation initiation, eukaryotic initiation factor, mRNA, open reading frame, untranslated region.*

a lack of ternary complex and a consequential down regulation of global translation.

Finding the initiation codon

The majority of translation initiation occurs by a scanning method[6], in which the small subunit of the ribosome is brought to the 5′ 7-methylguanosine (m^7G) capped-end of the mRNA by interactions with multiple eIFs. It then proceeds to migrate in a 3′ direction through the 5′ untranslated region (UTR) until a suitable initiation codon is encountered, thus defining the open reading frame (ORF). In prokaryotes, the scanning of the 30S small ribosomal subunit is halted by complementary base pairing between the mRNA and the 16S ribosomal RNA (rRNA) (Figure 1a). This 'Shine–Dalgarno' sequence[7] is located ten nucleotides upstream of the initiation codon, positioning the initiation codon in the P-site of the ribosome. This leads to recruitment of the 50S large ribosomal subunit, thus facilitating elongation.

Such an arrangement is not thought to occur in eukaryotes, although there is evidence that complementarity of the equivalent 18S rRNA with some mRNAs can enhance initiation[8,9]. Instead, the predominant model for eukaryotic scanning is that the initiation factors eIF1 and eIF1A 'read' the codons as they move through the ribosome and cause conformational changes upon encountering an AUG in a favourable consensus[10,11]. This consensus was postulated when relatively few mRNAs had been sequenced, where 699 mRNAs were examined. The frequency of bases surrounding the initiation codon was determined, with the highest frequencies deemed to be most favourable for initiation[12]. This Kozak consensus of GCC(A/G)CCAUGG (Figure 1b) was confirmed with mutagenesis experiments[13], recapitulated by others[14,15], where the G at +4 (the A of the AUG triplet being designated as +1) and the purine at −3 are central in determining initiation efficiency.

Multiple initiation sites in translation control

Whereas the above elegant model holds true in many cases, since its initial publication, multiple exceptions have been found. First, growth in sequence databases has yielded many thousands more mRNAs across species to consider. As such, AUG initiation codons have been identified which do not conform to the established consensus, and are considerably 'weaker' at initiating translation. Coupled with this is the discovery that many mRNAs have further AUG codons within their 5′ UTRs, which are either in- or out-of-frame with the annotated initiation codon[16]. Moreover, there are a growing number of ORFs which begin at a non-AUG initiation codon, albeit with reduced efficiency compared with an AUG codon. The use of out-of-frame initiation codons upstream of the physiological ORF (hence termed uORFs) to capture ribosomes and therefore reduce levels of translation at the physiological or annotated ORF (aORF) is one way in which mRNAs specifically control their own translation. mRNAs such as GCN4 in yeast and activating transcription factor 4 (ATF4) in mammals use

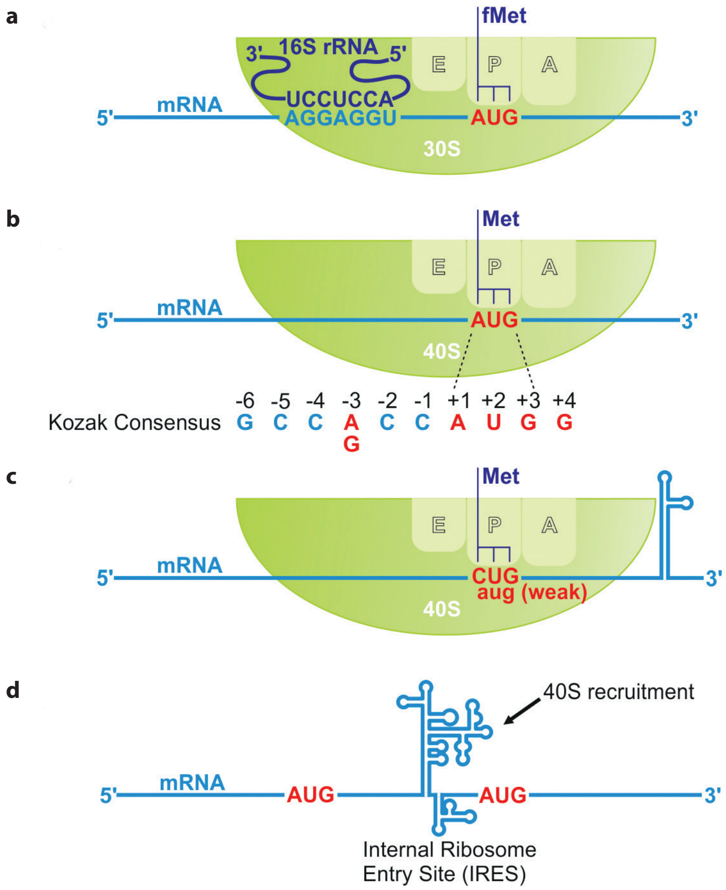

Figure 1. Mechanisms of initiation codon selection. **(a)** In prokaryotes, complementary base pairing between the 16S rRNA of the 30S small ribosomal subunit and the mRNA results in positioning of the initiation codon in the peptidyl-tRNA (P) site of the ribosome. This allows subsequent recruitment of the 50S large ribosomal subunit, and successive rounds of elongation beginning with the first formylmethionine (fMet). Charged tRNAs arrive in the aminoacyl-tRNA (A) site of the ribosome, with spent tRNAs from previous rounds of elongation being discarded via the exit (E) site. **(b)** Eukaryotic initiation codons exist within a consensus sequence that was first determined by Marilyn Kozak[12]. The purine at −3 and the G at +4 are particularly important in determining selection, with variants of this optimal sequence less efficient at initiating translation. **(c)** Translation at non-AUG codons (e.g. CUG) and AUGs in a weak Kozak consensus (shown as aug) may be enhanced if the RNA sequence downstream of the footprint of the 40S ribosomal subunit is capable of forming a stable secondary structure. Such structures could possibly arrest the scanning ribosome and create conditions favourable to initiation. **(d)** Secondary structures can form within some mRNAs that are able to recruit ribosomes directly, without the need for scanning from the 5′ cap-structure. These internal ribosome entry sites (IRESs) may allow potential upstream initiation codons to be bypassed.

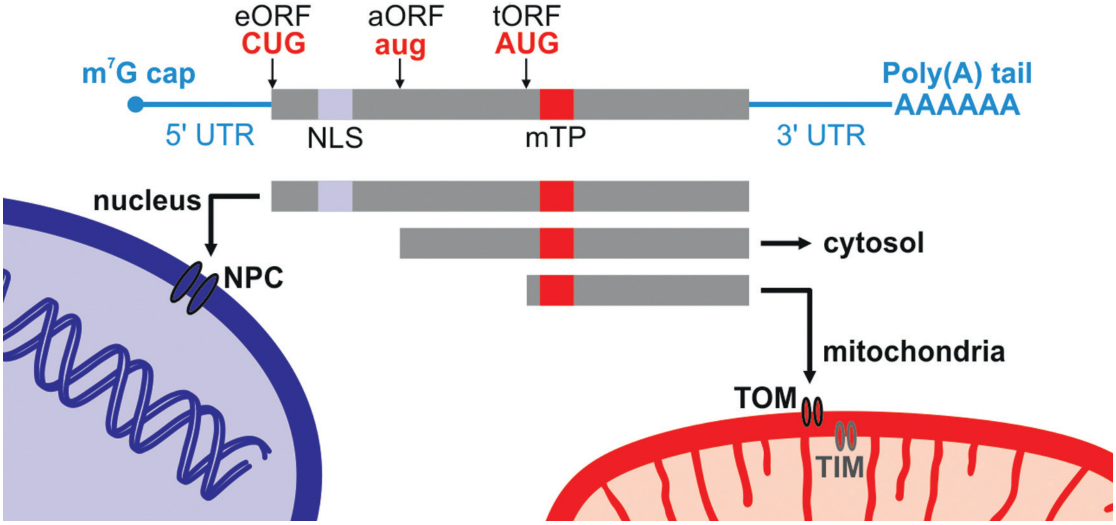

Figure 2. Potential consequences of alternative initiation. A theoretical protein with three potential in-frame initiation codons is illustrated. An extended ORF (eORF), translated from a non-AUG codon, results in production of a protein which contains a nuclear localization signal (NLS). This would be recognized by importins and translocated to the nucleus via the nuclear pore complex (NPC). The annotated form (aORF) translated from a weak AUG resides in the cytosol. A shorter truncated form (tORF), translated from an internal AUG contains a functional mitochondrial targeting peptide (mTP). Although the same sequence is present in the aORF form, it would not be detected by the mitochondrial targeting machinery due to it being further within the protein. Therefore only the tORF form would be recognized by the translocases of the outer and inner mitochondrial membranes (TOM and TIM), and transported to the mitochondrial matrix.

uORFs in a mechanism that controls the efficiency of translation of the physiological ORF[17–19]. For the remainder of this article, we concentrate on other instances, known as 'leaky scanning', where in-frame initiation codons (both AUG and non-AUG) lead to synthesis of proteins with differing N-termini, and the consequences of this phenomenon for the proteome.

Controlling fidelity of initiation codon selection

eIF1 and eIF1A play a pivotal role in recognition of AUG codons, and several studies have examined the consequences of manipulating the levels of these factors[20]. As expected, an increase leads to more stringent selection of those codons which adhere to the Kozak consensus, whereas knockdown leads to greater selection of AUGs in a weaker context and enhanced translation from non-AUG codons. Non-AUG initiation codons are usually 'near-cognate' meaning one base different from an AUG triplet (CUG, GUG, ACG, etc.), indicating that near-fidelity is still required for selection.

Moreover, eIF2A[21] and eIF2D[22] have been proposed as factors that aid in the selection of non-AUG codons by virtue of their ability to bring elongator tRNAs to an initiating ribosome. Whether these act in a failsafe mechanism to provide some, albeit limited, translation of mRNAs under conditions where eIF2 is phosphorylated is still a topic under discussion[23]. Unfortunately, determining what particular amino acid is incorporated at a non-AUG initiation codon (e.g. CUG decodes as leucine) is problematic as aminopeptidase enzymes remove the first one or two amino acids from the nascent peptide, therefore they cannot easily be identified by N-terminal sequencing[24].

In addition to the *trans*-acting factors contributing to recognition of initiation codons, the mRNAs themselves may aid the selection process by forming secondary structures which are inhibitory to scanning. The leading edge of the ribosome protrudes approximately 15 nucleotides beyond the initiation codon, and RNA hairpins at this position have been found to enhance translation from weak AUG or non-AUG codons[25] (Figure 1c). These 'buffers' possibly pause the scanning ribosome and enhance the likelihood that the changes brought about by eIF1 and eIF1A upon initiation codon recognition are favoured and hence initiation is undertaken.

More comprehensive secondary structures could act as internal ribosome entry sites (IRESs, Figure 1d), first found in picornaviral RNAs before their discovery in some cellular mRNAs[26]. These recruit ribosomes directly to internal sequences rather than via cap-dependent scanning. Although the majority do this to maintain translation under stress conditions at the usual annotated initiation codon, there are examples where an IRES can be used to translate from an internal initiation codon, thus producing an N-terminally truncated form of the protein[27,28].

Large-scale identification of novel initiation codons

The first documented non-AUG codon was an ACG codon in Sendai virus[29], and it was soon realized that such initiation codons existed in mammalian mRNAs. Case-by-case identifications slowly increased the number found, with fewer than 50 found by the mid-2000s. Given the huge expansion in gene databases by this time, attempts were thus made to examine the presence of RNA hairpins that could enhance initiation from weak initiation codons[30], to propose a consensus sequence for non-AUGs[31], and to also search for homology between *in silico* translated 5′ UTRs which may indicate that the sequence is in fact protein-coding[32,33]. A crucial innovation in the field has been the advent of next-generation sequencing technologies. For the transcription community, sequencing and quantification of mRNAs by RNA-Seq permits large-scale identification of splice variants and changes in promoter usage by mapping multiple short sequence reads back on to a genome. Within translation, this technology was initially used to sequence mRNA fragments that are protected by the ribosome during elongation. This method, termed Ribo-Seq, revealed hitherto undiscovered ORFs in many mRNAs and undoubtedly became one of the most significant advances in our understanding of the breadth of mRNA translation[34–37].

This technique has since been developed to use harringtonine or emetine, which prevent an elongation-ready ribosome from leaving the initiation codon, allowing mapping of translation initiation sites. Findings from a study in murine embryonic stem cells[36] revealed that established models of translation initiation of AUG usage, with one initiation site per mRNA, were perhaps not the rule but the exception. For example, less than half of the initiation events mapped actually used an AUG codon, ~60% of mRNAs had more than one initiation site and many transcripts contained uORFs, either separate to, or out-of-frame with and overlapping, the annotated ORF. Of particular relevance to this article was the discovery that ~15% of the transcripts had initiation sites that either extended or truncated the aORF, suggesting that many proteins have N-terminal variants. Moreover, advances in capture and sequencing of N-terminal peptides have confirmed the presence of multiple initiation sites in mRNAs within the proteome[38], thus proving that non-canonical initiation is a much more extensive phenomenon than previously acknowledged.

Why have multiple initiation codons?

Now that we realize that non-canonical and multiple initiation is a widespread process within gene expression, we should examine the biological reasons for such events. As mentioned above, one hypothesis is that non-AUG initiation may be a failsafe mechanism to ensure that limited translation of certain mRNAs is maintained during cellular stress[23]. Of more interest, however, is the fact that extensions and truncations may result in functionally distinct isoforms being produced. For example, alternatively initiated isoforms of the two-pore domain potassium (K2P) channels TREK1 and TREK2 show altered biophysical properties and ion selectivity[39,40], and our own work has determined that isoforms of the initiation factors eIF4GI and eIF4GII have differing activities in driving translation[41,42].

Another consequence of alternative initiation is represented in Figure 2, whereby three forms of a hypothetical protein have different subcellular locations depending on the initiation codon used. In our extended form (eORF), which uses a weaker non-AUG, a nuclear localization signal (NLS) is present, which results in translocation to the nucleus. The annotated form (aORF), translated from an AUG in a weak context, lacks the NLS, resulting in a cytosolic form. Such an example of altered localization has been determined for the BAG-1 protein[43]. As the first two initiation codons in our hypothetical protein were relatively weak, it is likely that the scanning ribosome is able to skip them and instead 'leak' to the third possible initiation codon creating a truncated form (tORF), which contains a functional mitochondrial targeting peptide (mTP). Such localization signals (also true of the ER signal peptide) are only recognized by the targeting machinery when presented at the very N-terminus of the protein, so would ordinarily be buried within the aORF. Such tORFs may exist in several mitochondrially localized proteins[44].

Leaky protein isoforms and expansion of the proteome

Until recently, the usage of alternative initiation codons to generate novel protein isoforms was overlooked compared with mechanisms such as alternative splicing. The use of Ribo-Seq and N-terminal peptide sequencing in identifying these events is clearly an exciting development and, as the technology becomes refined, cheaper and therefore more accessible to all, then our understanding of initiation codon selection will only improve further. What was once a trickle of knowledge about leaky protein isoforms is now becoming quite a substantial flood. ■

References

1. Beadle, G.W. and Tatum, E.L. (1941) Proc. Natl. Acad. Sci. U.S.A. **27**, 499–506
2. Hershey, J.W., Sonenberg, N. and Mathews, M.B. (2012) Cold Spring Harb. Perspect. Biol. **4**, a011528
3. Hinnebusch, A.G. (2011) Microbiol. Mol. Biol. Rev. **75**, 434–467
4. Sonenberg, N. and Hinnebusch, A.G. (2009) Cell **136**, 731–745
5. Wek, R.C., Jiang, H.Y. and Anthony, T.G. (2006) Biochem. Soc. Trans. **34**, 7–11
6. Kozak, M. (1978) Cell **15**, 1109–1123
7. Shine, J. and Dalgarno, L. (1974) Proc. Natl. Acad. Sci. U.S.A. **71**, 1342–1346
8. Chappell, S.A., Edelman, G.M. and Mauro, V.P. (2000) Proc. Natl. Acad. Sci. U.S.A. **97**, 1536–1541
9. Tranque, P., Hu, M.C., Edelman, G.M. and Mauro, V.P. (1998) Proc. Natl. Acad. Sci. U.S.A. **95**, 12238–12243
10. Lorsch, J.R. and Dever, T.E. (2010) J. Biol. Chem. **285**, 21203–21207
11. Passmore, L.A., Schmeing, T.M., Maag, D. et al. (2007) Mol. Cell **26**, 41–50
12. Kozak, M. (1987) Nucleic Acids Res. **15**, 8125–8148
13. Kozak, M. (1986) Cell **44**, 283–292
14. Ivanov, I.P., Loughran, G., Sachs, M.S. and Atkins, J.F. (2010) Proc. Natl. Acad. Sci. U.S.A. **107**, 18056–18060
15. Stewart, J.D., Cowan, J.L., Perry, L.S., Coldwell, M.J. and Proud, C.G. (2015) Biochem. J., doi:10.1042/BJ20141453
16. Peri, S. and Pandey, A. (2001) Trends Genet. **17**, 685–687
17. Barbosa, C., Peixeiro, I. and Romao, L. (2013) PLoS Genet. **9**, e1003529
18. Hinnebusch, A.G. (1993) Mol. Microbiol. **10**, 215–223
19. Hinnebusch, A.G. (1997) J. Biol. Chem. **272**, 21661–21664
20. Mitchell, S.F. and Lorsch, J.R. (2008) J. Biol. Chem. **283**, 27345–27349
21. Komar, A.A., Gross, S.R., Barth-Baus, D. et al. (2005) J. Biol. Chem. **280**, 15601–15611
22. Dmitriev, S.E., Terenin, I.M., Andreev, D.E. et al. (2010) J. Biol. Chem. **285**, 26779–26787
23. Starck, S.R., Jiang, V., Pavon-Eternod, M. et al. (2012) Science **336**, 1719–1723
24. Meinnel, T. and Giglione, C. (2008) Proteomics **8**, 626–649
25. Kozak, M. (1990) Proc. Natl. Acad. Sci. U.S.A. **87**, 8301–8305
26. Jackson, R.J. (2013) Cold Spring Harb. Perspect. Biol. **5**, a011569
27. Coldwell, M.J., deSchoolmeester, M.L., Fraser, G.A., Pickering, B.M., Packham, G. and Willis, A.E. (2001) Oncogene **20**, 4095–4100
28. Tinton, S.A., Schepens, B., Bruynooghe, Y., Beyaert, R. and Cornelis, S. (2005) Biochem. J. **385**, 155–163
29. Curran, J. and Kolakofsky, D. (1988) EMBO J. **7**, 245–251
30. Kochetov, A.V., Palyanov, A., Titov, I.I., Grigorovich, D., Sarai, A. and Kolchanov, N.A. (2007) BMC Bioinformatics **8**, 318
31. Wegrzyn, J.L., Drudge, T.M., Valafar, F. and Hook, V. (2008) BMC Bioinformatics **9**, 232
32. Ivanov, I.P., Firth, A.E., Michel, A.M., Atkins, J.F. and Baranov, P.V. (2011) Nucleic Acids Res. **39**, 4220–4234
33. Lin, M.F., Jungreis, I. and Kellis, M. (2011) Bioinformatics **27**, i275–i282
34. Ingolia, N.T. (2010) Methods Enzymol. **470**, 119–142
35. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. and Weissman, J.S. (2009) Science **324**, 218–223
36. Ingolia, N.T., Lareau, L.F. and Weissman, J.S. (2011) Cell **147**, 789–802
37. Lee, S., Liu, B., Huang, S.X., Shen, B. and Qian, S.B. (2012) Proc. Natl. Acad. Sci. U.S.A.**109**, E2424–E2432
38. Van Damme, P., Gawron, D., Van Criekinge, W. and Menschaert, G. (2014) Mol. Cell. Proteomics **13**, 1245–1261
39. Simkin, D., Cavanaugh, E.J. and Kim, D. (2008) J. Physiol. **586**, 5651–5663
40. Thomas, D., Plant, L.D., Wilkens, C.M., McCrossan, Z.A. and Goldstein, S.A. (2008) Neuron **58**, 859–870
41. Coldwell, M.J. and Morley, S.J. (2006) Mol. Cell. Biol. **26**, 8448–8460
42. Coldwell, M.J., Sack, U., Cowan, J.L. et al. (2012) Biochem. J. **448**, 1–11
43. Packham, G., Brimmell, M. and Cleveland, J.L. (1997) Biochem. J. **328**, 807–813
44. Kazak, L., Reyes, A., Duncan, A.L. et al. (2013) Nucleic Acids Res. **41**, 2354–2369

Mark Coldwell is an Associate Professor in Biochemistry, working in the Centre for Biological Sciences at the University of Southampton. He received his PhD in Biochemistry from the University of Leicester in 2001 for work on internal ribosome entry sites in cellular mRNAs, and has continued to work on the regulation of translation initiation in mammalian models ever since. Following postdoctoral work at the University of Sussex working on the eIF4G eukaryotic initiation factor scaffolds, Mark has been building a research group at the University of Southampton since 2008. The group particularly concentrates on the discovery of alternative initiation codons and other translational control mechanisms. email: M.Coldwell@soton.ac.uk

Joanne Cowan is a BBSRC-funded Postdoctoral Research Fellow in the Coldwell group. She received her PhD from the University of Sussex, working on translational control during proteasome inhibition and myogenic differentiation. She then worked as a medical writer before resuming her laboratory research career at Southampton. Her current work involves examining alternative initiation events in human genes and the resultant subcellular localizations of the novel isoforms. email: J.L.Cowan@soton.ac.uk