Trends in Biochemical Sciences

CellPress
REVIEWS

Opinion

Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes

Greco Hernández,[1,*] Vincent G. Osnaya,[1] and Xochitl Pérez-Martínez[2]

Selection of the translation initiation site (TIS) is a crucial step during translation. In the 1980s Marylin Kozak performed key studies on vertebrate mRNAs to characterize the optimal TIS consensus sequence, the Kozak motif. Within this motif, conservation of nucleotides in crucial positions, namely a purine at −3 and a G at +4 (where the A of the AUG is numbered +1), is essential for TIS recognition. Ever since its characterization the Kozak motif has been regarded as the optimal sequence to initiate translation in all eukaryotes. We revisit here published *in silico* data on TIS consensus sequences, as well as experimental studies from diverse eukaryotic lineages, and propose that, while the −3A/G position is universally conserved, the remaining variability of the consensus sequences enables their classification as optimal, strong, and moderate TIS sequences.

Highlights

The Kozak sequence has been characterized as a conserved TIS in eukaryotes.

Sequencing of mRNAs from diverse species coupled to functional studies will continue to contribute to understanding the impact of sequence context on TISs with cognate or non-cognate initiation codons.

The 3D resolution of both translational initiation factors and ribosomes on mRNA will enable better understanding of their conservation, as well as the conservation of the TIS sequences.

Ongoing sequencing projects, such as the Earth BioGenome Project, will help to determine the conservation of the Kozak sequence between eukaryotes such as vertebrates, plants, and fungi.

Conservation and Variability of the Kozak Motif

Translation, namely the synthesis of proteins by the ribosome and translation factors using mRNA as the template, is a fundamental process for all forms of life because proteins catalyze the vast majority of reactions sustaining life and play structural, transport, and regulatory roles in all living organisms. In eukaryotes, regulating gene expression at the level of translation is crucial for tissues or developmental stages where transcription is quiescent or limited, as well as for the differential spatial distribution of proteins in cells and tissues [1]. Regulating translation also gives cells the potential to elicit rapid and reversible responses to sudden environmental changes and stresses without invoking mRNA transcription, processing, or transport [1]. Thus, translational control plays a significant role in determining both protein abundance and proteome composition, and, accordingly, a myriad of mechanisms to regulate translation have evolved throughout eukaryotic evolution [2,3]. Recognition of the mRNA AUG **codon** (see Glossary) that initiates translation, termed the **translation initiation site** (TIS), is subject to tight control because this step establishes the correct open reading frame for mRNA decoding (Box 1).

In the decades following 1980, Marylin Kozak analyzed the frequency of nucleotides around the TIS of vertebrate **coding sequences** (CDSs) in mRNAs [4–6]. She established that the sequence GCC*R*CCAUGG (where R at −3, in italic font, is A or G; and the AUG **initiation codon** is underlined) is both the consensus sequence flanking the TIS and the one most optimal for translation in vertebrate lineages; this sequence is termed the **Kozak motif** [7]. Interestingly, many new TIS sequences of phylogenetically distant, non-vertebrate phyla have been analyzed with the advent of the genomic era and high-throughput technologies, as well as with the development of different systems in which to study translation (discussed below). Even so, there is a lack of a comprehensive view of TIS consensus sequences across eukaryotes. We revisit here published data with a primary focus on vertebrates, unicellular fungi, insects, flowering land plants, and some protists to explore these data and evaluate whether the Kozak motif is also the TIS consensus sequence in different taxa. Remarkably, we observe that the −3R position is universally conserved among all eukaryotes scrutinized here, but that there is variation in the +4 position, particularly in single-celled fungi and some protists. Surprisingly, we observed that A/C at the −2 position is universally conserved as well. We also noticed that there are differences between eukaryotic lineages, and therefore propose that TIS consensus sequences can be categorized as Kozak optimal, strong, and moderate motifs.

The Kozak Motif: Optimal, Strong, and Moderate Beginnings

During translation initiation, the small (40S) ribosomal subunit binds to the mRNA at the **5′-untranslated region** (5′-UTR) and **scans** in the 5′ to 3′ direction to reach the CDS start codon, most usually an

[1]Translation and Cancer Laboratory, Unit of Biomedical Research on Cancer, National Institute of Cancer (Instituto Nacional de Cancerología, INCan), 22 San Fernando Avenue, Tlalpan, 14080 Mexico City, Mexico

[2]Department of Molecular Genetics, Cell Physiology Institute (Instituto de Fisiología Celular), Universidad Nacional Autónoma de México (UNAM), 04510 Mexico City, Mexico

*Correspondence: ghernandezr@incan.edu.mx

Box 1. Translation Initiation in Eukaryotes

Translational control primarily occurs at the initiation step in which mRNA is recruited to the ribosome (reviewed in [1,25,92,93] and references therein). Translation begins with the recognition of the cap structure (m^7GpppN, where N is any nucleotide) located at the 5′ end of the mRNA by eIF4E. A 40S ribosomal subunit bound to eIF1, eIF1A, eIF3, and eIF5 then promotes recruitment of a **ternary complex** (TC), consisting of eIF2 attached to GTP and an initiator Met-tRNA$_i^{Met}$, to form a 43S **preinitiation complex** (PIC). This step positions the initiator Met-tRNA$_i^{Met}$ in the peptidyl (P) site of the ribosome. In a parallel set of interactions, the scaffold protein eIF4G interacts with the poly(A)-binding protein (PABP), the RNA helicase eIF4A, the ribosome-bound eIF3, and the cap-bound eIF4E to coordinate recruitment of the 43S PIC to the mRNA 5′-UTR. eIF4A unwinds secondary structures of 5′-UTR, allowing the 43S PIC to scan base-by-base the 5′-UTR in the search for an AUG initiation codon to start translation [1,25,92,93].

eIF1 and eIF1A drive selection of the correct start codon. They cooperatively promote the adoption of an **open conformation** of the 43S PIC that is compatible with scanning. Such conformation features the Met-tRNA$_i^{Met}$ loosely engaged in the P-site, and allows eIF1 and eIF1A to discriminate against codon–**anticodon** mismatches. 5′-UTR scanning proceeds until an authentic initiator codon, most often an AUG codon, is reached and establishes correct codon–anticodon basepairing. This event arrests scanning and promotes transition of the 43S PIC to a **closed conformation** in which Met-tRNA$_i^{Met}$ and eIF1A become tightly positioned within the P-site. Together, these actions result in the formation of a 48S PIC. Then, GTP–eIF5B promotes the release of eIF1 and eIF5B, facilitating binding of a 60S subunit to the 48S PIC to assemble an 80S initiation complex [1,25,92,93].

AUG [8,9]. The scanning model predicts that translation initiates at the AUG codon closest to the 5′ end of the mRNA, which is termed the 'first-AUG rule' [8,9]. Landmark studies to understand the role of the flanking sequence in TIS recognition by the ribosome led Kozak to analyze 153 cellular and viral mRNAs from vertebrates, observing that 'the sequences flanking the TIS are not random' [4]. She initially found that *R*NNAUG*G* (where N is any base) is the consensus sequence surrounding the AUG initiator codon, and that the two positions which show the most significant conservation, −3R and +4G, function to generate the most efficient initiation signal [4]. She later analyzed 211 cellular mRNAs, mostly from vertebrates but also from some insects, sea urchins, and flowering land plants, and observed that CC*R*CCAUG*G* was the consensus sequence around the TIS [5]. Kozak then analyzed 699 vertebrate mRNAs, and from this study an expanded consensus for the TIS context emerged, namely (GCC)GCC*R*CCAUG*G* in which 97% of mRNAs have a purine, most often an A (61%), at the −3 position [6].

Site-directed mutagenesis experiments to analyze the functional contribution of every nucleotide at positions −1 to −10 and the G at +4 position showed that translation decreases five to tenfold when either −3A or the +4G is replaced by C or U, and 20-fold when pyrimidines are substituted at both of those positions ([4,7,8] and references therein, [10]). These studies also demonstrated that the −3R position plays a key role in TIS recognition and that, in its absence, the +4G position also significantly contributes to promoting translation [7,10,11]. Lütcke *et al.* also demonstrated that, in a vertebrate system, the translational efficiency of an mRNA was 100, 85, 61, and 38% for A, G, U, and C at position −3, respectively [12]. By contrast, positions −1C, −2C, −4C, −5C, and −6G slightly contributed to promoting translation only in the absence of both −3R and +4G. Nucleotides in positions −7 to −10 showed essentially no influence on translation [7,10,11]. Likewise, the +5 and +6 positions also had no impact on translation [11].

Additional experiments showed that changing −3R and/or +4G caused some 40S ribosomal subunits to bypass the first AUG and instead initiate at the next AUG [7,13] in a mechanism termed 'leaky scanning' [8,14]. However, because recognizing the TIS position is crucial, ribosomes may initiate at AUG codons even if they lie in a weak context (i.e., lacking the −3R and/or +4G). Indeed, a −3A is usually sufficient for ribosomes to recognize the first AUG irrespective of the other nucleotides in the context [8,14]. It has also been shown that stem-loop structures ~14 nt downstream of a suboptimal AUG context improve its recognition by the 40S ribosome, possibly by slowing the scanning process. This facilitation is most significant when the ribosome stalls directly over the AUG [15].

Glossary

Aminoacyl (A) site: the site on the 40S ribosomal subunit that holds the incoming aminoacyl-tRNA.
Anticodon: the triplet of a tRNA that is complementary to an mRNA codon.
Biosphere: layer of Earth where life exists. The biosphere is one of the four layers that surround the Earth, together with the lithosphere (soil and rock), hydrosphere (water), and atmosphere (air), and the biosphere is the sum of all the ecosystems and communities on a global scale.
Coding sequences (CDSs): usually start with an AUG codon and end with any of the stop codons. In a few examples, non-cognate codons such as GUG, CUG, ACG, or AUU are used to start the coding sequences.
Closed conformation: a ribosomal state of codon–anticodon base pairing that results in displacement of eIF1s from the P-site; also called the 'in' conformation.
Codon: a sequence of three RNA nucleotides that encodes a specific amino acid (or stop signal) during protein synthesis.
Eukaryotic initiation factors (eIFs): soluble proteins that drive the initiation phase of eukaryotic translation, in other words the mRNA recruitment to the ribosome and the further formation of a ribosomal preinitiation complex at an initiation codon.
Initiation codon: mRNA triplet that opens a reading frame for protein synthesis. The majority of peptides initiate with an AUG codon which encodes methionine. Some open reading frames start with non-AUG codons (also termed 'near-cognate'), which may be GUG, CUG, ACG, and, very rarely, AUU.
Kozak motif: a consensus sequence surrounding the mRNA translation initiation site (TIS, usually the codon AUG) that was discovered by Marilyn Kozak in the 1980s in vertebrate mRNAs.
Microbial dark matter: microbial world composed of genes, genomes, bacteria, archaea, protists, and viruses of unknown identity, as well as their processes and communities. The term is by analogy to the dark matter of the cosmos studied by astronomers.

Table 1. Interactions between Factors, Ribosome, and mRNA during TIS Recognition

Species	Protein	Amino acid	Contacts	Refs
Saccharomyces cerevisiae (budding yeast)	eIF1	Arg36	Codon/anticodon duplex	[19]
Yeast	eIF1	Asp71, Glu73, Glu76	D stem backbone of Met-tRNA$_i$	[19]
Tetrahymena thermophila	eIF1	Arg26, Arg27, Gly28, Arg29, and Lys30	Codon/anticodon duplex	[20]
Oryctolagus cuniculus (Rabbit)	eIF1	Arg38–Lys42	+4 of mRNA, codon/anticodon duplex	[21]
Rabbit	eIF1	Pro77–Gly80	D stem of tRNA$_i^{Met}$	[21]
Tetrahymena thermophila	eIF1A	Asn43, Arg45, Trp69, Lys87	rRNA	[22]
Yeast	eIF1A	Gly8, Gly9, Lys10	Codon/anticodon duplex	[19]
Yeast	eIF1A	Lys16	+5 of mRNA	[19]
Yeast	eIF1A	Trp70	AUG and +4 of mRNA	[19]
Rabbit	eIF1A	Amino terminal	+4 and antisense loop of tRNA$_i$	[21]
Yeast	eIF2α	Arg55	−3 of mRNA	[18,19]
Yeast	eIF2α	Arg57	−2 of mRNA	[18,19]
Rabbit	eIF2α		−3 of mRNA	[17,24]
Rabbit	rpS26e	Val83	−3 of mRNA	[21]
Rabbit	rpS5		−3 and −4 of mRNA	[17,24]
Rabbit	rpS15		+4 and +5 of mRNA	[17,24,26]

Overall, these experiments showed that the context sequence strongly influences TIS recognition by the ribosome in vertebrate mRNAs. Mainly, positions −3R (most often A) and +4G are both the most conserved nucleotides and exert the most critical influence on translational efficiency. Thus, the Kozak motif GCC*R*CCAUG*G* was established as the optimal context for TIS recognition in vertebrate mRNAs [8]. Further, depending on the presence of the two crucial nucleotides (i.e., −3*R* and +4*G*), sequences surrounding the TIS have been classified as 'optimal', GCC*R*CCAUG*G*; 'strong', NNN*R*NNAUG*G* (only the two important nucleotides are present); 'adequate', NNN*R*NNAUG(A/C/U) or NNN(C/U)NNAUG*G* (only one of these nucleotides is present); and 'weak', NNN(C/U)NNAUG(A/C/U) (any sequence lacking both key nucleotides) Kozak motifs [16].

Consistent with the importance of a purine at the −3 and +4 positions, Pisarev *et al.* showed an interaction between these nucleotides and the translational machinery [17]. They observed interactions between −3A and the **eukaryotic initiation factor** (eIF) 2α subunit, and of +4G with ribosomal protein S9 (rpS9) and 18S ribosomal RNA (rRNA). Further biochemical, genetic, and structural studies in yeast, rabbit, and *Tetrahymena thermophila* have recently led to a full understanding of the mRNA TIS recognition by the 40S ribosome **peptidyl (P) site**, eIF1, eIF1A, eIF2α, and rpS26e at the atomic level (Table 1 and Box 2) [18–27].

Open conformation: a complete ribosomal complex in which initiation factors eIF1, eIF1A, eIF2, and eIF4F, together with the 40S subunit, establish a state that promotes scanning; also called the 'out' conformation.
Peptidyl (P) site: ribosomal tRNA-binding site that holds the peptidyl-tRNA.
Preinitiation complex (PIC): a macromolecular complex containing the 40S small ribosomal subunit (bound to eIF1A, eIF1, and eIF3) and the ternary complex. This complex requires the cooperative action of eIF4F to attach to the mRNA.
Scanning: ribosomal movement in a 5′ to 3′ direction along the 5′-UTR of an mRNA to unwind secondary structures (promoted by RNA helicases) to reach an initiation codon.
Ternary complex (TC): molecular complex comprising eIF2, a molecule of GTP, and an initiator Met-tRNA (Met-tRNA$_i$). It binds the 40S ribosomal subunit to promote 43S complex formation during translation initiation.
Translation initiation site (TIS): mRNA triplet that opens the reading frame for protein synthesis. TIS triplets are most frequently AUG (encoding methionine), but other non-cognate codons such as GUG (encoding valine) and CUG (encoding leucine) may sometimes act as a TIS.
5′-Untranslated region (5′-UTR): the mRNA 5′-UTR is located upstream of the AUG start codon. It does not contain a protein coding sequence but is crucial for regulating translation by a variety of mechanisms.

The TIS Consensus Context in Disparate Lineages

Sequence Analyses of TIS Consensus Sequences

More recently, the TIS context of CDSs from many species belonging to phylogenetically distant taxa, namely single-celled fungi, insects, vertebrate, flowering land plants, and some protists, have been analyzed *in silico* by different research groups to identify TIS consensus sequences (Tables S1–S5 in the supplemental information online). The universe of published consensus motifs prompted us to determine their distribution across the tree of life [28,29], and to ask whether or not the optimal vertebrate Kozak motif GCC*R*CCAUG*G* is also the TIS consensus sequence in non-vertebrate species.

In analyzing these data, different authors have used various criteria to select a 'consensus' sequence. Many of them used the most common procedure in the field, namely the Cavener consensus rule [30,31]. According to this rule, a particular nucleotide at a specific position is given a consensus status indicated by a capital letter if its frequency is greater than 50% and greater than twice the frequency of the second most frequent nucleotide at that position. When no single nucleotide satisfies these criteria, a pair of bases are assigned co-consensus status and indicated by capital letters if the sum of their frequencies is greater than 75%. If no single nucleotide or pair of nucleotides meet these criteria, the most frequent nucleotide is denoted by a lower-case letter.

Thus, by using the Cavener consensus rule, we recalculated the published consensus sequences of those reports that did not use it. We observed only slight variations between the recalculated sequences and the original sequences (Tables S1–S5). We then categorized all consensus sequences according to the Kozak motifs defined above [16]. We choose to use the term 'moderate' instead of 'adequate' for those TIS contexts that possess only one of the key nucleotides. We did not recalculate the sequences of Nakagawa *at al.* [32], Pesole *et al.* [33], and Acevedo *et al.* [34] because the raw data were not provided. Figure 1 (Key Figure) depicts a global view of the distribution of the consensus sequences surrounding the CDSs TIS across the tree of life, according to the current classification of eukaryotes [28,29].

Over the years, many studies have corroborated that the optimal Kozak motif GCC*R*CCAUG*G*, or a slight variation of it, GCM*R*NCAUG*G* (where M is an A or C), is actually the consensus sequences around the TIS in vertebrate (phylum Chordata) mRNAs, demonstrating the highly predictive power of Kozak's findings (Table S1) [31,32,35–37]. However, some of the species analyzed, such as cattle, pig, red junglefowl, and the fish *Danio rerio*, possess strong motifs NNN*R*NCAUG*G* as the consensus

Box 2. Recognition of the AUG Start Codon

To date, recognition of the correct AUG start codon by initiation factors in eukaryotes is well known at the ultrastructural level (Figure I) [9,18–22,92]. During the initiation of translation, eIF1 and eIF1A attach to the 40S subunit near the P-site and the **aminoacyl (A) site**, respectively. For mRNA scanning to proceed, eIF1 and the eIF1A C-terminal tail (CTT) act together to discriminate against non-AUG codons. When non-AUG codons are detected, the Arg38–Lys42 stretch of rabbit eIF1 spatially interferes with the anticodon stem-loop of $tRNA_i^{Met}$ and nucleotide +4 of mRNA [21]. The Pro77–Gly80 region of rabbit eIF1 also obstructs interaction with the D stem of $tRNA_i^{Met}$ [21]. These steric clashes destabilize mismatches between codon–anticodon duplexes and result in the open conformation that promotes scanning (Figure IA). In brief, the open conformation does not allow full accommodation of $tRNA_i^{Met}$ in the P-site owing to clashes with eIF1 and eIF1A-CTT [18–25,92].

Upon recognition of an AUG codon by correct basepairing with the anticodon of Met-$tRNA_i^{Met}$, yeast eIF1A N-terminal tail (NTT) residues Gly8, Gly9, and Lys10 interact with the codon–anticodon duplex and stabilize it, while yeast eIF1A Lys16 and Trp70 contact nucleotides +4 and +5 of the mRNA [19,21]. These interactions inhibit scanning, stabilize codon–anticodon duplex formation, and promote a transition to the closed conformation that displays tight accommodation of $tRNA_i^{Met}$ within the P-site. AUG recognition also drives conversion of GTP–eIF2 to GDP–eIF2, liberating phosphate (Pi), as well as dissociation of eIF1, GDP–eIF2, and eIF1A-CTT from the P-site. Transition to the closed conformation (Figure IB) brings rabbit Arg55 and Arg57 of eIF2α D1 loop in contact with mRNA nucleotides −2 and −3, and Val83 of rabbit rpS26e to contact nucleotide −3. These interactions promote the initiation of translation from an AUG in an appropriate context [18–25,92].

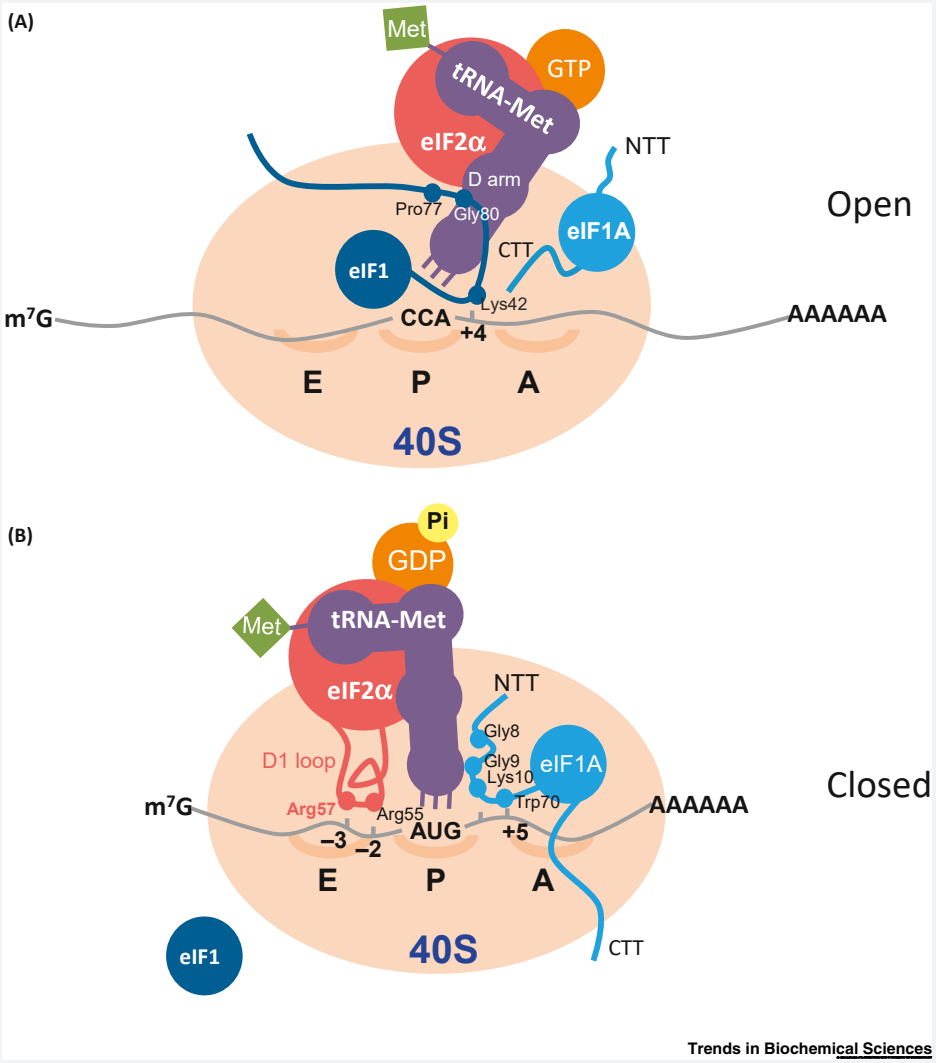

Figure I. Recognition of the mRNA AUG Start Codon by the Ribosome.

(A) Ribosome open conformation during scanning. (B) Once a TIS is recognized, that ribosome changes to a closed conformation that is incompatible with the scanning process.

context instead of an optimal motif [32,37]. Indeed, the −6G, −5C, and −1C positions are relatively well conserved across vertebrate consensus sequences (Table S1 and Table 2). Further, tunicates (phylum Chordata), the closest living relatives of vertebrates, occupy the most basal position in the chordate phylogeny and have been extensively studied to understand the evolutionary origin of vertebrates [38]. Within this lineage, the sea squirt *Ciona intestinalis* possesses a moderate Kozak motif with a K at the +4 position (where K is a G or U) as the preferred TIS consensus sequence [32] (Table S1). This observation could suggest that the optimal Kozak motif for vertebrates appeared at the emergence of the chordates. It would be exiting to study whether or not this motif is also the TIS consensus motif in other tunicate species. We also noticed that the –2M position is highly conserved among the consensus sequences of the vertebrates and the tunicate studied (Table 2).

Key Figure

The Consensus Translation Initiation Site (TIS) Kozak Motif Varies among Different Eukaryotes

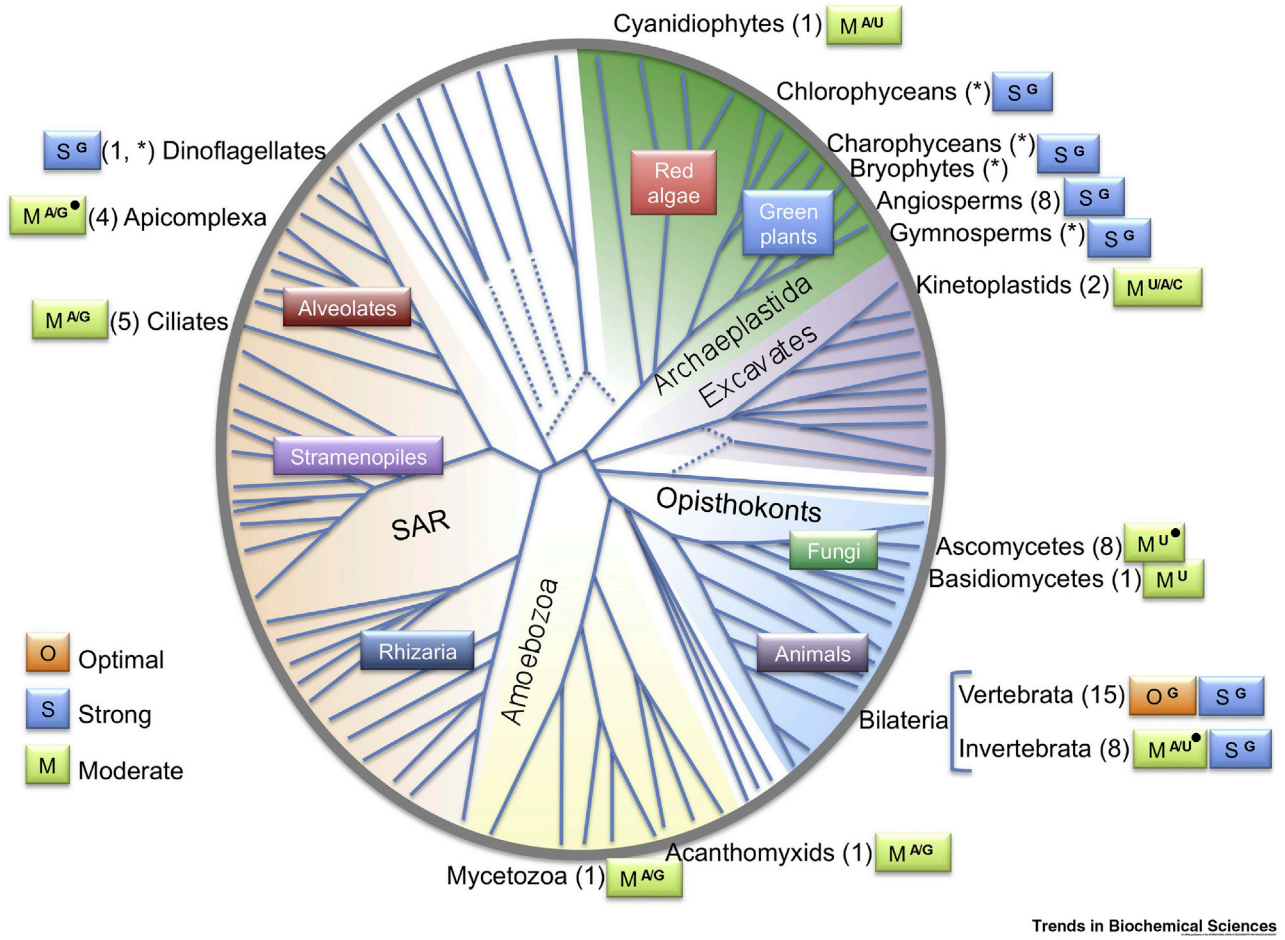

Figure 1. The A or G at the −3 position is conserved in all lineages studied. Other positions (including the functionally key +4 nucleotide) of the consensus Kozak motif vary among different eukaryotes. As a preferred sequence, the optimal Kozak motif is restricted to vertebrates and some angiosperm (monocotyledon) plants. Strong Kozak motifs are present in some vertebrates, flowering land plants (supergroup Archaeoplastida), and Dinoflagellates (supergroup Alveolates). Moderate Kozak motifs are the prevalent TIS consensus in fungi, invertebrates, apicomplexans, ciliates, mycetozoans, and acanthomyxidans. Numbers in brackets indicate the number of species analyzed in each lineage, an asterisk (*) indicates cases where the number was not specified by the authors. (●) Indicates lineages that generally use moderate Kozak motifs but contain a few species that use strong Kozak motifs (Tables S1–S5 for details). The preferentially used nucleotide(s) at the +4 position are indicated as superscripts. Classification of eukaryotes is according to [28,29]. Kozak motifs are defined according to [16]: O, optimal GCC*R*CCAUG*G*; S, strong NNN*R*NNAUG*G*; M, moderate NNN*R*NNAUG(A/C/U) or NNN(C/U)NNAUG*G*. Abbreviations: N, any nucleotide; R, purine (A or G); SAR, stramenopiles, alveolates, and rhizaria. Scheme adapted from Burky [28] with permission from Cold Spring Harbor Laboratory Press.

Among invertebrate species, several insects (phylum Arthropoda) have been analyzed (Table S2) [31,32,34,37,39]. Whereas the fruit fly *Drosophila melanogaster* possesses a strong motif [31,34,37], species from different orders, including *Apis mellifera*, *Bombyx mori*, and *Tribolium castaneum* on the one hand, and *Anopheles gambiae* on the other, contain moderate consensus motifs of the type NNM*R*MMAUGK and NNN*R*NNAUGW (where W is A or U), respectively [32,39]. Further,

invertebrate from other phyla, such as the roundworm *Caenorhabditis elegans* (phylum Nematoda) and the blood fluke *Schistosoma japonicum* (phylum Platyhelminthes), also possess moderate motifs as TIS consensus flanking sequences [31,32]. More specifically, *A. mellifera*, *B. mori*, *T. castaneum*, and *S. japonicum* have moderate consensus motifs with a K at the +4 position [31,32,39]. Interestingly, the −4M and −2M positions appear to be relatively well conserved in the consensus sequences of the invertebrates mentioned here (Table 2).

Within the supergroup Archaeplastida (Plantae), several species of green plants (land plants and green algae) have been analyzed (Table S3). They possess strong Kozak motifs as the consensus TIS context [31–33,36,40–43]. The frequencies of TISs containing guanine at positions −3 and +4 were observed to be different between monocotyledon and dicotyledon mRNAs. Whereas monocot mRNAs contain higher frequencies of −3G/+4G nucleotides, −3A/+4G are more frequent in dicot mRNAs. It should be pointed out that the 5′-UTRs of monocot mRNAs are GC-rich, whereas those of dicots are AU-rich sequences [31–33,40,41,43]. Moreover, we have noticed that, among the consensus sequences of the plant species here scrutinized, the −2M position is well conserved (Table 2). A single species of red algae has also been analyzed, namely *Cyanidioschyzn merolae* [32], which possesses a moderate consensus motif.

All fungi (supergroup Opisthokonts) analyzed have been unicellular species, including *Saccharomyces cerevisiae*, *Aspergillus fumigatus*, *Deberyomyces hansenii*, *Yarrowia lipolytica*, *Eremothecium gossypii*, *Kluyveromyces lacti*, *Candida glabrata*, *Schizosaccharomyces pombe* (phylum Ascomycetes), and *Cryptococcus neoformans* (phylum Basidiomycetes) (Table S4). In contrast to the animals and plants mentioned above, these unicellular fungi mainly have moderate Kozak motifs with a bias towards U at the +4 position [31–33,44,45] (Figure 1). Moreover, the −2M and −4M positions also appear to be relatively well conserved (Table 2). It would be interesting to analyze whether or not this is also the case in multicellular fungi.

Table S5 shows protist species from diverse phyla of the supergroups Amebozoa and Excavates, and the group Alveolates. Among them, different species of dinoflagellates (Alveolates), such as *Symbiodinium kawatii*, use strong Kozak motifs as TIS consensus sequences [46,47]. By contrast,

Table 2. Most Conserved Positions among TIS Consensus Sequences from Coding Sequences in Different Groups of Eukaryotes[a]

Group	Sequence[b]							
	−6	−5	−4	−3	−2	−1	AUG	+4
Vertebrates	G	C		*R*	M	C	———	*G*
Tunicata			M	*A*	M		———	G/U
Invertebrates			M	*R*	M		———	G/A/U
Green plants				*R*	M		———	*G*
Red algae				*A*	M	C	———	A/U
Fungi			M	*R*	M		———	U
Amoebozoa (Mycetozoa, Acanthomyxids)				*R*	M		———	R
Alveolates (Dinoflagellates, Apicomplexa, and Ciliates)								
Excavata (Kinetoplastids)				*R*	M		———	C/A/U
Universally conserved				*R*	M		———	

[a]Owing to its key functional importance, the +4 position is also shown despite its lack of conservation.
[b]Key: ———, AUG; M is A or C; R is A or G.

apicomplexan species (Alveolates) such as *Theileria* spp., *Cryptosporidim parvum*, and *Plasmodium falciparum* [32,48], the kinetoplastids (Excavata) *Leishmania major* and *Trypanosome brucei* [32], the mycetozoan (Amebozoa) *Dictyostelium* spp. and *Acanthamoeba castellanii* [31,32,49], and the ciliates *Paramecium* spp., *Tetrahymena* spp., *Euplotes* spp., *Oxytricha* spp., and *Stylonyclia lemnae* [49], prefer moderate Kozak motifs with different nucleotides (namely G, A, U, or C) at the +4 position as the consensus TIS context. Interestingly, although the −2M position is well conserved in all protist consensus sequences included here, M at the −4 position is also conserved among apicomplexan, dinoflagellate, and mycetozoan species (Table 2).

Recently, a myriad of giant, novel viruses have been discovered that contain substantial sets of translational factors, including eIF1 and eIF2α [50,51]. It would be exciting to investigate whether or not those factors are involved in the host TIS context recognition.

Experimentally Tested TIS Consensus Sequences

Despite the high value of *in silico* analyses, the functional importance of most TIS consensus sequences remains to be experimentally evaluated. Indeed, only some sequences from a few species have been tested in the laboratory (Table S6). As mentioned, the optimal Kozak motif has experimentally been shown *in vitro* to be the best sequence for promoting translation in vertebrates, in particular, rat and rabbit [7,10,11]. Despite these findings, for other vertebrates such as the model organism zebrafish (*Danio rerio*), the optimal Kozak sequence was found to be a poor predictor of translation efficiency [37]. In this case, the investigators found that, although the optimal Kozak sequence efficiently promoted translation *in vivo*, it is neither the most frequent nor the most efficient sequence for initiating translation. Instead, the most frequent sequence is almost twice as efficient at promoting translation as the optimal Kozak motif [37]. Thus, the capacity of the sequences to promote translation may vary among different species, and therefore should be experimentally tested in the specific organism of interest.

Experimental studies of invertebrate species are scarce. Only two insects have been analyzed, namely *D. melanogaster* [34] and *B. mori* [39]. According to what has been observed *in silico*, strong Kozak motifs have the most influence on promoting translation in *Drosophila* [34]. In the case of *B. mori*, a difference was observed between the *in silico*-obtained consensus TIS sequence [32,39] and the most efficient sequence determined experimentally [39]. In particular, the +4 position appears to be different. This apparent discrepancy might be explained by the significant dissimilarity between the number of analyzed sequences in both approaches, namely 14 sequences *in vitro* [39] versus 50 [39] and 875 sequences [32] *in silico*.

Regarding plants, the influence of the TIS context on translation initiation has been experimentally examined in some flowering species (Table S6). *In vitro* translation assays observed no differences in translation efficiency in the wheat system when −3A was changed for G, U, or C [12]. Other *in vitro* assays with 21 sequences did find that −3A/G and +4G are the most efficient nucleotides for translation initiation [4]. Later, *in vivo* experiments in tobacco (16 sequences) [43], *Arabidopsis thaliana* and *Oryza sativa* (64 sequences) [52], and *Picea abies* and *Zea mays* (both 16 sequences) [43] have shown that −3A/G and +4G indeed confer the best translational efficiency.

Among fungi, only *S. cerevisiae* has been experimentally studied, although some discrepancies have been observed. Early studies by Looman and Kuivenhoven (1993) of 47 sequences showed that this yeast prefers the moderate Kozak motif AAUUANNAUGUCU for optimal mRNA translation [53], in agreement with different *in silico* analyses of genomic sequences that also established a moderate sequence with a +4U as the TIS consensus context [31,32,44,45]. Among these studies, Nakagawa *et al*. [32] analyzed 5980 CDSs. By contrast, Pesole *et al*. inspected 1378 CDSs *in silico* [33], obtaining the strong Kozak motif CAMMAUGG as the TIS consensus context. In agreement with this result, ribosome profiling experiments (also termed Ribo-Seq) enabled the analysis of 1735 mRNAs and observed that both core nucleotides (−3A and +4G) are widely used in yeast [54]. The different results between the genomic and the Ribo-Seq observations might reflect incorrect annotation of the main

CDS AUGs in the earliest studies because we now know that there may be considerable use of upstream AUGs or non-AUG codons of CDSs. We could also explain these differences by the fact that the genomic studies analyzed all annotated genes, whereas the second approach investigated the set of mRNAs being translated at a specific moment under specific growth conditions. More studies will be necessary to better understand the optimal TIS Kozak motif in *S. cerevisiae*. So far, apart from *S. cerevisiae*, no other unicellular organisms have been experimentally analyzed in this regard.

Conservation and Variability among Eukaryotes

Overall, the *in silico* and experimental studies show that, across taxa, the most highly conserved position in the consensus TIS context and the nucleotide that influences translation the most is the purine at position −3 (most frequently A). The conservation of the initiation factors involved in TIS recognition, together with the fact that all taxa here examined possess identical preferences for the crucial −3 position, is an indication of the mechanistic similarities in translation initiation.

Interestingly, the −2A/C position appeared to be the second most conserved position in those lineages scrutinized here (Table 2). This observation is in agreement with early studies by Nakagawa *et al.* who reported a strong bias towards A/C at position −2 in 47 genomes of diverse species of animals, fungi, plants, and protists [32]. They found that genes with higher expression levels showed stronger signals, suggesting that nucleotides at these positions are involved in the regulation of translation initiation [32]. Accordingly, a contact between yeast eIF2α and the −2 position has been observed during TIS recognition by the translation machinery [18,19]. Position −2A/C is present in chordates, the invertebrate, plants, single-celled fungi, and protist TIS consensus sequences. This observation means that the translation machinery that recognizes the start codon is highly conserved. In the line with this idea, sequence comparisons of the translational machinery factors driving recognition of the TIS consensus sequence, namely eIF1, eIF1A, eIF2α, and ribosomal protein S26 (rpS26), showed no significant differences in the amino acids involved in TIS and Kozak motif recognition among the species reviewed here (Figures S1–S4). Indeed, the percentage of identity among these factors is in the same range of identity of other initiation factors such as eIF4E and eIF4A from the same lineages (Figure S5), suggesting that the variations in those factors causing a context preference are subtle, or that more mechanistic details of context recognition remain to be uncovered.

Early studies of genomic sequences of 47 species from different taxa by Nakagawa *et al.* [32] showed that the diversity of TIS consensus motifs can be arranged into two distinct general patterns, namely GCCGCCAUG and AAAAAAAUG. By reviewing a much broader spectrum of species from different taxa, we have observed significant variability of Kozak motifs among different species (Figure 1 and Table 2). The experimental and *in silico* studies described here agree with *in silico* analyses that optimal and strong Kozak motifs (i.e., with a +4G) appear to be the best sequences to promote translation in vertebrates [5,6,31–33,35]. In *Drosophila*, land plants, and dinoflagellates, strong Kozak motifs (with a +4G) are also the preferred sequence surrounding the CDS TIS [31–34,36,40–43,46,47]. Interestingly, the consensus sequence flanking the AUG initiator codon appears to have diverged in other eukaryotes. In particularly, nucleotides at positions −6, −5, −4, and −1, as well as the key position +4, in single-celled fungi (Ascomycetes) [31–33,44,45] and some protists (for instance some ciliates, kinetoplastids, apicomplexans, and mycetozoans [31,32,48,49] show a preference for moderate Kozak motifs with a bias towards a +4U (Ascomycetes) and +4A/C/U for the aforementioned protists. Thus, the optimal Kozak motif does not appear to be the optimal sequence to initiate translation in all lineages. Finally, we noticed that –3U/C and –2U/G are universally absent nucleotides in the consensus TIS motif for CDSs. Because 18S rRNA is also involved in TIS recognition, differences in this rRNA among species might also contribute to differential TIS context recognition. It would be interesting to experimentally study the sequences flanking the TISs in a larger number of species of different phyla to better understand the distribution of Kozak motifs across the tree of eukaryotes.

Strong Initiation at Non-AUG Initiator Codons

Despite the predominant use of the AUG codon to start CDS translation across eukaryotic lineages, decoding in some mRNAs initiates at codons different from AUG. Although translation starting at non-AUG codons is much less efficient than canonical initiation, CUG, GUG, ACG, and UUG are the most commonly used near-cognate codons [55]; codons AUU, AUC, AUA, and AGG may also be used, albeit very rarely [55].

Recent advances in ribosome profiling aiming to precisely map TISs genome-wide in different systems, particularly human, mouse, and *S. saccharomyces*, have revealed thousands of alternative translation initiation events at non-AUG codons [56,57]. These approaches are based on the action of antibiotics such as lactimidomycin, harringtonine, puromycin, and cycloheximide which block ribosome activity at different steps of translation. However, there is ongoing debate on the value of some of these data because it remains unknown what proportion of these ribosomal footprints are due to experimental artifacts and how many represent actual translation initiation events [55,58,59]. It is noteworthy that a computational analysis found a weak correlation between context strength and TIS efficiency for the non-AUGs codons obtained in several Ribo-Seq datasets [60]. By contrast, the use of genetically engineered reporters in mammals indicates that translation from non-AUGs is more dependent on their TIS context than is translation from AUG codons, although sequence context affects each non-AUG start codon differently [61].

We reviewed studies that experimentally addressed the translation of non-AUG codons in their natural contexts (Table S7). In vertebrate mRNAs (*H. sapiens*, *M. musculus*, *O. cuniculus*, *Ratus norvegicus*, and the monkey *Chlorocebus aethiops*) there is a sharp bias to use either strong or moderate Kozak motifs around non-AUG TISs to achieve the most efficient translation [13,62–77]. This context usage contrasts with the AUG TIS of vertebrate CDSs which prevalently prefers an optimal Kozak motif as the consensus context [5,6,32,33,36]. The flowering land plants *N. plumbaginifolia*, *Orychophragmus violaceus* [69], wheat [70], and *A. thaliana* [71], as well as the fungi *N. crassa* [78] and *S. cerevisiae* [72,74], also prefer either optimal or strong Kozak motifs flanking non-AUGs TISs in different transcripts. In the case of the ascomycetes *S. cerevisiae* [73] and *Candida albicans* [75], the galactokinase and CARP2A mRNAs, respectively, possess moderate contexts with an A at the +4 position. Among protists, *P. falciparum* uses the moderate Kozak context and start codon UUUUUUUUAGG in aldolase mRNA (interestingly, UAG is a stop codon in the canonical genetic code) [76].

In Table S8 we have reviewed the influence of context on the translation of non-AUG TISs in experimentally tested human [79] and *S. cerevisiae* [80–82] transcripts. Although human non-AUGs prefer strong Kozak motifs (NNN*R*NNAUG*G*) for most efficient translation, *S. cerevisiae* may prefer moderate (NNN*A*NNAUGA or NNN*A*NNAUGR) Kozak motifs for efficient translation. Again, more sequences and species should be tested to understand in more depth the influence of context on non-AUG TIS recognition.

Concluding Remarks

The evolution of eukaryotes towards a myriad of different lineages has resulted in large-scale changes and significant innovations at many levels. At the molecular level, this includes the translation process. Emerging evidence suggests that the fundamental mechanisms of translation are well conserved in all eukaryotes, even though the initiation step has undergone substantial increases in sophistication upon the emergence of eukaryotes [2,83,84]. Likewise, the studies reviewed here show that the TIS consensus context also has diverged to some extent during eukaryotic evolution, and that this diversity includes the crucial nucleotide at the +4 position; thus, some lineages appear to possess optimal Kozak consensus sequences, some have moderate to strong TIS consensus sequences, and others may utilize various TIS consensus sequences. Near-cognate start codon TISs from the few species analyzed to date, including vertebrates, show a strong preference for a strong, but not optimal, TIS consensus context. It is remarkable that, despite these changes, for both AUG and non-AUG initiator codons, A or G (most frequently A) at the −3 position is a universal feature of all taxa analyzed so

far. The second most conserved position across the lineages scrutinized here appears to be A/C at the −2 position, and U/C and U/G at the –3 and –2 positions are universally absent in the consensus Kozak motif for CDSs.

In the near future we believe that several recently launched and highly ambitious global sequencing initiatives will unlock a wealth of highly valuable novel information on the **biosphere**. These projects include the Earth BioGenome Project that aims to sequence the genomes of all of eukaryotic species on Earth over a period of 10 years [85], metagenomics projects that plan to publish the genomes of over 100 000 species of the **microbial dark matter** within the next 5 years [86–89], and projects that will extensively study the spectrum of viruses in thousands of geographically diverse samples to uncover the so-called 'virome of Earth' [90,91]. Undoubtedly, the field of translation will benefit from these ground-breaking projects (see Outstanding Questions).

Outstanding Questions

Is the Kozak motif of monophyletic origin, meaning that it appeared in the last ancestor of all eukaryotes, or it did appear several times during eukaryotic evolution?

From the data discussed herein, the −3 position is the most conserved base in the TIS sequence. Is the purine at the −3 position conserved across all eukaryotes?

After the publication of global sequencing initiatives such as the Earth BioGenome Project, will we find further diversity in TIS consensus sequences across eukaryotes?

It is well documented that translational control plays a key role in cancer onset and progression. Is the prevalence of translation of mRNAs containing specific Kozak motifs (i.e., weak, moderate, strong, or optimal) altered during cancer development?

Acknowledgments

G.H. was financed by internal funding of the National Institute for Cancer (INCan, Mexico). V.G.O. was supported by a Master in Sciences Scholarship from the Consejo Nacional de Ciencia y Tecnologia (CONACyT; grant 355715), Mexico. X.P-M. was supported by CONACyT (grant 284514) and the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT)–UNAM (IN209217). We thank Cold Spring Harbor Laboratory Press for permission to adapt Figure 1 (permission 186208/1156417P). We also appreciate valuable criticism and comments from anonymous reviewers and the editors that significantly improved the manuscript.

Supplemental Information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.tibs.2019.07.001.

References

1. Hershey, J.W.B. *et al.* (2018) Principles of translational control. *Cold Spring Harb. Perspect. Biol.* a032607
2. Hernández, G. *et al.* (2016) On the origin and early evolution of translation in eukaryotes. In *Evolution of the Protein Synthesis Machinery and Its Regulation* (Hernández, G. and Jagus, R. eds), pp. 81–108, Springer
3. Hernández, G. *et al.* (2010) Origins and evolution of the mechanisms regulating translation initiation in eukaryotes. *Trends Biochem. Sci.* 35, 63–73
4. Kozak, M. (1981) Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucleic Acids Res.* 9, 5233–5252
5. Kozak, M. (1984) Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.* 12, 857–872
6. Kozak, M. (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* 15, 8125–8148
7. Kozak, M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44, 283–292
8. Kozak, M. (2002) Pushing the limits of the scanning mechanism for initiation of translation. *Gene* 299, 1–34
9. Hinnebusch, A.G. (2014) The scanning mechanism of eukaryotic translation initiation. *Annu. Rev. Biochem.* 83, 779–812
10. Kozak, M. (1987) At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J. Mol. Biol.* 196, 947–950
11. Kozak, M. (1997) Recognition of AUG and alternative inititator codons is augmented by G in position +4 but is not generally affected by nucleotides in position +5 and + 6. *EMBO J.* 16, 2482–2492
12. Lütcke, H.A. *et al.* (1987) Selection of AUG initiation codons differs in plants and animals. *EMBO J.* 6, 43–48
13. Kozak, M. (1989) Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol. Cell. Biol.* 9, 5073–5080
14. Kozak, M. (1991) Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. Biol. Chem.* 266, 19867–19870
15. Kozak, M. (1990) Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc. Natl. Acad. Sci. U. S. A.* 87, 8301–8305
16. Meijer, H.A. and Thoma, A.A.M. (2002) Control of eukaryotic protein synthesis by upstream open reading frames in 5′-untranslated region of an mRNA. *Biochem. J.* 367, 1–11
17. Pisarev, A. *et al.* (2006) Specific functional interactions of nucleotides at key −3 and +4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. *Genes Dev.* 20, 624–636
18. Llacer, J.L. *et al.* (2015) Conformational differences between open and closed states of the eukaryotic translation initiation complex. *Mol. Cell* 59, 1–14
19. Hussain, T. *et al.* (2014) Structural changes enable start codon recognition by the eukaryotic translation initiation complex. *Cell* 159, 597–607
20. Rabl, J. *et al.* (2011) Crystal structure of the eukaryotic 40S ribosomal subunit in complex with initiation factor 1. *Science* 331, 730–736
21. Lomakin, I.B. and Steitz, T.A. (2013) The initiation of mammalian protein synthesis and mRNA scanning mechanism. *Nature* 500, 307–311

22. Weisser, M. *et al.* (2013) The crystal structure of the eukaryotic 40S ribosomal subunit in complex with eIF1 and eIF1A. *Nat. Struct. Mol. Biol.* 20, 1015–1017
23. Pestova, T.V. and Kolupaeva, V.G. (2002) The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. *Genes Dev.* 16, 2906–2922
24. Pisarev, A.V. *et al.* (2008) Ribosomal position and contacts of mRNA in eukaryotic translation initiation complexes. *EMBO J.* 27, 1609–1621
25. Lind, C. and Åqvist, J. (2016) Principles of start codon recognition in eukaryotic translation initiation. *Nucleic Acids Res.* 44, 8425–8432
26. Bulygin, K. *et al.* (2005) The first position of a codon placed in the A site of the human 80S ribosome contacts nucleotide C1696 of the 18S rRNA as well as proteins S2, S3, S3a, S30, and S15. *Biochemistry* 44, 2153–2162
27. Martin-Marcos, P. *et al.* (2011) Functional elements in initiation factors 1, 1A, and 2β discriminate against poor AUG context and non-AUG start codons. *Mol. Cell. Biol.* 31, 4814–4831
28. Burki, F. (2014) The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* 6, a016147
29. Adi, S.M. *et al.* (2019) Revisions to the classification, nomenclature, and diversity of eularyotes. *J. Eukaryot. Microbiol.* 66, 4–119
30. Cavener, D.R. (1987) Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res.* 15, 1353–1361
31. Cavener, D.R. and Ray, S.C. (1991) Eukaryotic start and stop translation sites. *Nucleic Acids Res.* 19, 3185–3192
32. Nakagawa, S. *et al.* (2008) Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res.* 36, 861–871
33. Pesole, G. *et al.* (2000) Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. *Gene* 261, 85–91
34. Acevedo, J.M. *et al.* (2018) Changes in global translation elongation or initiation rates shape the proteome via the Kozak sequence. *Sci. Rep.* 8, 4018
35. Kochetov, A.V. *et al.* (1998) Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS Lett.* 440, 351–355
36. Rangan, L. *et al.* (2008) Analysis of context sequence surrounding translation initiation site from complete genome of model plants. *Mol. Biotechnol.* 39, 207–213
37. Grzegorski, S.J. *et al.* (2014) Natural variability of Kozak sequences correlates with function in a zebrafish model. *PLoS One* 9, e108475
38. Kourakis, M.J. and Smith, W.C. (2015) An organismal perspective on *C. intestinalis* development, origins and diversification. *eLife* 4, e06024
39. Tatematsu, K. *et al.* (2014) Effect of ATG initiation codon context motifs on the efficiency of translation of mRNA derived from exogenous genes in the transgenic silkworm, *Bombyx mori*. *Springerplus* 3, 136
40. Gupta, P. *et al.* (2016) Comparative analysis of contextual bias around the translation initiation sites in plant genomes. *J. Theoret. Biol.* 404, 303–311
41. Joshi, C.P. *et al.* (1997) Context sequences of translation initiation codon in plants. *Plant Mol. Biol.* 35, 993–1001
42. Kawaguchi, R. and Bailey-Serres, J. (2005) mRNA sequence features that contribute to translational regulation in *Arabidopsis*. *Nucleic Acids Res.* 33, 955–965
43. Lukaszewicz, M. *et al.* (2000) *In vivo* evaluation of the context sequence of the translation initiation codon in plants. *Plant Sci.* 154, 89–98
44. Hamilton, R. *et al.* (1987) Compilation and comparison of the sequence context around the AUG startcodons in *Saccharomyces cerevisiae* mRNAs. *Nucleic Acids Res.* 15, 3581–3593
45. Cigan, A.M. and Donahue, T.F. (1987) Sequence and structural features associated with translational initiator regions in yeast. *Gene* 59, 1–18
46. Zhang, H. *et al.* (2013) Proof that dinoflagellate spliced leader (DinoSL) is a useful hook for fishing dinoflagellate transcripts from mixed microbial samples: *Symbiodinium kawagutii* as a case study. *Protist* 164, 510–527
47. Bodył, A. and Mackiewicz, P. (2007) Analysis of the targeting sequences of an iron-containing superoxide dismutase (SOD) of the dinoflagellate *Lingulodinium polyedrum* suggests function in multiple cellular compartments. *Arch. Microbiol.* 187, 281–296
48. Saul, A. and Battistutta, D. (1990) Analysis of the sequences flanking the translational start sites of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* 42, 55–62
49. Yamauchi, K. (1991) The sequence flanking translational initiation site in protozoa. *Nucleic Acids Res.* 19, 2715–2720
50. Abrahão, J. *et al.* (2018) Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nat. Comm.* 9, 749
51. Abrahão, J. *et al.* (2017) The analysis of translation-related gene set boosts debates around origin and evolution of mimiviruses. *PLoS Genet.* 13, e1006532
52. Sugio, T. *et al.* (2010) Effect of the sequence context of the AUG initiation codon on the rate of translation in dicotyledonous and monocotyledonous plant cells. *J. Biosci. Bioeng.* 109, 170–173
53. Looman, A.C. and Kuivenhoven, J.A. (1993) Influence of the three nucleotides upstream of the initiation codon on expression of the *Escherichia coli* lacZ gene in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 21, 4268–4271
54. Robbins-Pianka, A. *et al.* (2010) The mRNA landscape at yeast translation initiation sites. *Bioinformatics* 26, 2651–2655
55. Kearse, M.G. and Wilusz, J.E. (2017) Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev.* 31, 1717–1731
56. Ingolia, N.T. *et al.* (2018) Ribosome profiling: global views of translation. *Cold Spring Harb. Perspect. Biol.* 11, a032698
57. Ingolia, N.T. (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* 15, 205–213
58. Andreev, D.E. *et al.* (2017) Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res.* 45, 513–526
59. Guttman, M. *et al.* (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154, 240–251
60. Andreev, D.E. and Baranov, P.V. (2014) Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning. *BMC Bioinformatics* 15, 380
61. Diaz de Arce, A.J. *et al.* (2018) Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Res.* 46, 985–994
62. Florkiewicz, R.Z. and Sommer, A. (1989) Human basic fibroblast growth factor gene encodes four polypeptides: three initiate translation from non-AUG codons. *Proc. Natl. Acad. Sci. U. S. A.* 86, 3978–3981

63. Arnaud, E. *et al.* (1999) A new 34-kilodalton isoform of human fibroblast growth factor 2 is cap dependently synthesized by using a non-AUG start codon and behaves as a survival factor. *Mol. Cell. Biol.* 19, 505–514
64. Huez, I. *et al.* (2001) New vascular endothelial growth factor isoform generated by internal ribosome entry site-driven CUG translation initiation. *Mol. Endocrinol.* 15, 2197–2210
65. Fuxe, J. *et al.* (2000) Translation of p15.5INK4B, an N-terminally extended and fully active form of p15INK4B, is initiated from an upstream GUG codon. *Oncogene* 19, 1724–1728
66. Tikole, S. and Sankararamakrishnan, R. (2000) A survey of mRNA sequences with a non-AUG start codon in RefSeq database. *J. Biomol. Struct. Dyn.* 24, 33–42
67. Acland, P. *et al.* (1990) Subcellular fate of the int-2 oncoprotein is determined by choice of initiation codon. *Nature* 343, 662–665
68. Bruening, W. and Pelletier, J. (1996) A non-AUG translational initiation event generates novel WT1 isoforms. *J. Biol. Chem.* 271, 8646–8654
69. Gordon, K. *et al.* (1992) Efficient initiation of translation at non-AUG triplets in plant cells. *Plant J.* 2, 809–813
70. Peabody, D.S. (1989) Translation initiation at non-AUG triplets in mammalian cells. *J. Biol. Chem.* 264, 5031–5035
71. Christensen, A.C. *et al.* (2005) Dual-domain, dual-targeting organellar protein presequences in *Arabidopsis* can use non-AUG start codons. *Plant Cell* 17, 2805–2816
72. Clements, J.M. *et al.* (1988) Efficiency of translation initiation by non-AUG codons in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 8, 4533–4536
73. Zitomer, R.S. *et al.* (1984) *Saccharomyces cerevisiae* ribosomes recognize non-AUG initiation codons. *Mol. Cell. Biol.* 4, 1191–1197
74. Donahue, T.F. and Cigan, A.M. (1988) Genetic selection for mutations that reduce or abolish ribosomal recognition of the HIS4 translational initiator region. *Mol. Cell. Biol.* 8, 2955–2963
75. Abramczyk, D. *et al.* (2003) Non-AUG translation initiation of mRNA encoding acidic ribosomal P2A protein in *Candida albicans*. *Yeast* 20, 1045–1052
76. Ghersa, P. *et al.* (1990) Initiation of translation at a UAG stop codon in the aldolase gene of *Plasmodium falciparum*. *EMBO J.* 9, 1645–1649
77. Prats, H. *et al.* (1989) High molecular mass forms of basic fibroblast growth factor are initiated by alternative CUG codons. *Proc. Natl. Acad. Sci. U. S. A.* 86, 1836–1840
78. Wei, J. *et al.* (2013) The stringency of start codon selection in the filamentous fungus *Neurospora crassa*. *J. Biol. Chem.* 288, 9549–9562
79. Mehdi, H. *et al.* (1990) Initiation of translation at CUG, GUG, and ACG codons in mammalian cells. *Gene* 91, 173–178
80. Chen, S.J. *et al.* (2008) Translational efficiency of a non-AUG initiation codon is significantly affected by its sequence context in yeast. *J. Biol. Chem.* 283, 3173–3180
81. Chen, S.J. *et al.* (2009) Translational efficiency of redundant ACG initiator codons is enhanced by a favorable sequence context and remedial initiation. *J. Biol. Chem.* 248, 818–827
82. Chang, C.P. *et al.* (2010) A single sequence context cannot satisfy all non-AUG initiator codons in yeast. *BMC Microbiol.* 10, 188
83. Hernández, G. (2009) On the origin of the cap-dependent initiation of translation in eukaryotes. *Trends Biochem. Sci.* 34, 166–175
84. Hernández, G. *et al.* (2012) On the diversification of the translational apparatus across eukaryotes. *Comp. Funct. Genomics* 2012, 256848
85. Lewin, H.A. *et al.* (2018) Earth BioGenome Project: sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4325–4333
86. Parks, D.H. *et al.* (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2, 1533–1542
87. Solden, L. *et al.* (2016) The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Curr. Opin. Microbiol.* 31, 217–226
88. Hugenholtz, P. *et al.* (2016) Genome-based microbial taxonomy coming of age. *Cold Spring Harb. Perspect. Biol.* 8, a018085
89. Kyrpides, N.C. *et al.* (2016) Microbiome data science: understanding our microbial planet. *Trends Microbiol.* 24, 425–427
90. Paez-Espino, D. *et al.* (2016) Uncovering Earth's virome. *Nature* 25, 425–430
91. Khalil, J. *et al.* (2016) Updating strategies for isolating and discovering giant viruses. *Curr. Opin. Microbiol.* 31, 80–87
92. Hinnebusch, A.G. (2017) Structural insights into the mechanism of scanning and start codon recognition in eukaryotic translation initiation. *Trends Biochem. Sci.* 42, 589–611
93. Pelletier, J. and Sonenberg, N. (2019) The organizing principles of eukaryotic ribosome recruitment. *Annu. Rev. Biochem.* 88, 307–335