ISSN 0006-2979, Biochemistry (Moscow), 2023, Vol. 88, Suppl. 1, pp. S176-S191 © Pleiades Publishing, Ltd., 2023.
Russian Text © The Author(s), 2023, published in Uspekhi Biologicheskoi Khimii, 2023, Vol. 63, pp. 349-378.

REVIEW

The Influence of the Nucleotide Composition of Genes and Gene Regulatory Elements on the Efficiency of Protein Expression in *Escherichia coli*

Artur I. Zabolotskii[1,a]*, Stanislav V. Kozlovskiy[1], and Alexey G. Katrukha[1]

[1] *Faculty of Biology, Lomonosov Moscow State University, 119991 Moscow, Russia*
[a] *e-mail: zabolotscky.artur@yandex.ru*

Received May 25, 2022
Revised June 23, 2022
Accepted June 29, 2022

Abstract—Recombinant proteins expressed in *Escherichia coli* are widely used in biochemical research and industrial processes. At the same time, achieving higher protein expression levels and correct protein folding still remains the key problem, since optimization of nutrient media, growth conditions, and methods for induction of protein synthesis do not always lead to the desired result. Often, low protein expression is determined by the sequences of the expressed genes and their regulatory regions. The genetic code is degenerated; 18 out of 20 amino acids are encoded by more than one codon. Choosing between synonymous codons in the coding sequence can significantly affect the level of protein expression and protein folding due to the influence of the gene nucleotide composition on the probability of formation of secondary mRNA structures that affect the ribosome binding at the translation initiation phase, as well as the ribosome movement along the mRNA during elongation, which, in turn, influences the mRNA degradation and the folding of the nascent protein. The nucleotide composition of the mRNA untranslated regions, in particular the promoter and Shine–Dalgarno sequences, also affects the efficiency of mRNA transcription, translation, and degradation. In this review, we describe the genetic principles that determine the efficiency of protein production in *Escherichia coli*.

DOI: 10.1134/S0006297923140109

Keywords: recombinant proteins, codon composition, codon optimization, *Escherichia coli*, expression activation, solubility

INTRODUCTION

Recombinant proteins are widely used in food, biomedical and other fields of biotechnology. *Escherichia coli* remains the most popular organism among a large number of species used for the expression of non-glycosylated recombinant proteins [1, 2]. The advantages of this microorganism include a comprehensive understanding of its metabolic processes, high growth rate, relatively low price and availability of cultivation media, possibility of scaling up, and existence of a large number of *E. coli* strains, expression vectors, and genetic engineering tools [1, 2].

Increasing protein expression is one of the key problems in both industrial protein production and laboratory research, and significant efforts are directed to the development of methods for its optimization. Various approaches, such as selection of the optimal strains, expression vectors, and expression conditions (nutrition media, cultivation methods, protein synthesis induction) are already widely used for increasing the level of protein expression in *E. coli* [3]. However, the problem of low protein yield is often associated with the non-optimal nucleotide sequences of the expressed genes and their regulatory regions.

Both coding and noncoding gene sequences contain elements that influence protein folding and protein yield at all stages of protein production. The effects of many factors involved in the control of expression are often

Abbreviations: CAI, codon adaptation index; CDS, coding sequence; GFP, green fluorescent protein; nTE, normalized translational efficiency; RF, release factor; SD, Shine–Dalgarno (sequence); tAI, tRNA adaptation index; TIR, translation initiation region; UTR, untranslated region.

* To whom correspondence should be addressed.

interrelated, which hinders our comprehension of their action mechanisms individually or in their combination.

Nevertheless, many years of studies have yielded a noticeable progress in the understanding of the influence of nucleotide sequences on the protein production. In this review, we present the current ideas on the influence of nucleotide composition of genes and their regulatory sequences on the protein production at different stages of expression in *E. coli*.

INFLUENCE OF GENE NUCLEOTIDE COMPOSITION ON THE PROTEIN EXPRESSION EFFICIENCY

Transcription. *Promoter and adjacent regions.* The first step in the protein synthesis is gene transcription by the DNA-dependent RNA polymerase (RNAP). The initiation rate and the efficiency of mRNA synthesis are determined by the gene promoter sequence, a region upstream of the gene coding sequence (CDS) that provides the binding of the corresponding proteins (RNA polymerase, transcription factors) for the transcription initiation. There are two main types of promoters used for the recombinant protein expression: inducible (regulated promoters activated only in response to a specific stimulus) and constitutive promoters (permanently active unregulated promoters). Most often, the promoter strength is evaluated experimentally from the relative level of mRNA or reporter protein (most often, green fluorescent protein, GFP) produced by expression under control of this promoter [4, 5]. Promoter sequences and their strength (determined experimentally) can be found in the BIOFAB (International Open Facility Advancing Biotechnology; http://parts.igem.org/Collections/BioFAB), Anderson promoter library (http://parts.igem.org/Promoters/Catalog/Anderson), and other databases.

The most common approach for developing efficient promoters is creation of libraries with randomized promoter sequences followed by the analysis of the reporter gene expression under control of these promoters. For example, see the reports on the randomization of the −10/−35 sequence of the constitutive *Ptrc* promoter [6] or the −17/+3 sequence of the inducible *T7* promoter [7, 8]. Although most of the key principles of transcription initiation are known, the models for promoter strength prediction based on its nucleotide sequence are still at the stage of development. However, it was experimentally demonstrated that using a strong promoter almost always results in the reproducible increase in the gene transcription and, consequently, in the protein expression [6].

The rate of RNAP-catalyzed transcription. Another factor affecting the rate of transcription is the content of hydrogen bonds, or the percent content of GC pairs (GC%), in the DNA sequence from the transcription initiation region to codon ~15 of the CDS. The GC% content affects the energy required for the RNA polymerase to melt the DNA duplex and, hence, determines the rate of the early transcription stages. It was shown that in *E. coli* cells (but not in eukaryotes), gene sequences with a lower GC% content in the 5' region are translated more efficiently [9].

Stability and toxicity of mRNA. *mRNA toxicity*. Expression of heterologous proteins often slows down the cell growth, which is usually associated with the toxicity of expressed proteins or high metabolic burden [10]. However, it was found that some expressed heterologous mRNAs were toxic for *E. coli* cells. Thus, transcription of some synonymous variants of the GFP gene in *E. coli* inhibited cell growth independently on the gene translation [11]. I was suggested that the mechanism of this effect, which is currently actively studied, is associated with the toxic effect of specific mRNA secondary structures.

mRNA degradation. The coupling of transcription and translation in bacteria results in mRNA coverage by the ribosomes and various protein factors already at the stage of mRNA synthesis, which protects mRNA against degradation by RNases. A high density of ribosomes on the actively translated mRNAs can efficiently prevent the action of endoribonucleases [12]. However, comparing the stability of bacterial and eukaryotic mRNAs reveals that bacterial mRNAs exist for a relatively short time. The half-life of most bacterial mRNAs varies from 40 seconds to 60 minutes, whereas eukaryotic mRNAs can exist for up to several days [13]. This is believed to be associated with the absence in prokaryotes of specific mechanisms and proteins protecting mRNAs against degradation (e.g., those similar to polyA-binding protein in eukaryotes) [14]. The 5'- and 3'-untranslated regions (UTRs) of mRNAs, which lack bound ribosomes, are especially sensitive to the endonuclease activity and determine the stability of mRNAs in *E. coli*.

Thus, the 5'-UTR of the outer membrane protein (OmpA) mRNA forms a secondary structure that ensures the stability and increases fourfold the half-life of this mRNA [15, 16]. It is believed that the presence of highly structured 5' regions inhibits the binding of endoribonucleases (e.g., RNase E) to the 5'-terminal regions of mRNAs unprotected by proteins [17]. However, it is difficult to evaluate the role of the 5'-terminal region stability in the efficiency of the mRNA expression, because the nucleotide composition of this region also plays a key role in the extremely important process of translation initiation [18-20], for which presence of the secondary structures is utterly unfavorable.

Changes in the 3'-UTR can also prolong the half-life of mRNA and elevate the protein expression level. Thus, a replacement and/or shortening of the mRNA 3'-end enhanced the mRNA stability and increased the amount of the produced protein [21]. This effect can

be due to: (i) a decrease in the formation on the mRNA 3'-end of secondary structures susceptible to the action of RNases specific toward the double-stranded RNA [22]; (ii) disruption of the interaction between the 3'- and 5'-ends leading to the suppression of the mRNA degradation by RNases [23]; (iii) the presence of specific motifs (AU-rich motifs [24], RNase recognition sites, sRNA-binding sites, etc. [22]).

Translation. *Translation initiation. Efficiency of ribosome binding.* In *E. coli*, the rate of translation and the protein yield are largely determined by the translation initiation; therefore, this stage is the most important target in the optimization of protein expression [25, 26]. The translation initiation rate depends on multiple factors, one of them being the affinity of the 16S rRNA of the 30S ribosomal subunit to the Shine–Dalgarno (SD) sequence located in the mRNA translation initiation region (TIR). It was shown that the energy of binding between the 16S rRNA and SD sequence is in a good correlation with the observed levels of protein expression [6, 27, 28].

Despite this correlation, all attempts to develop universal SD sequences providing efficient synthesis of any CDS have been unsuccessful. A standard SD sequence that successfully initiates translation of one sequence, might not work with another sequence, thus requiring the development of the optimal SD sequence for each individual CDS [6, 27, 29]. Most likely, this is due to the fact that the nucleotide composition of the SD sequence region might affect the formation of the mRNA secondary structures in the TIR, which, in their turn, determine the protein expression level. The replacement of the SD sequence can lead to the formation on the mRNA 5'-end of the secondary structures decreasing the efficiency of translation initiation and protein expression. For this reason, the SD sequence is typically optimized for each gene.

Bicistronic design (BCD) is an example of the SD sequence optimization without disturbance of the secondary structure of the mRNA 5'-end [6]. This method is based on translational coupling of two fused sequences, wherein the translation of the downstream CDS depends on the translation of the upstream cistron and supposedly results from the reinitiation of translation by the same ribosomes[30, 31]. The optimal structure of the TIR is provided by the SD sequence and the short CDS of the first cistron, which prevents the influence of the mRNA secondary structures on the downstream CDS, thus allowing selection of optimal SD variant. In theory, such separation of the second CDS should simplify the development of standardized SD sequences for the efficient translation initiation of any CDS [6, 27, 29].

TIR secondary structure. Numerous studies have shown that the level of protein expression in *E. coli* is most strongly affected by the formation of stable secondary structures in the mRNA 5'-terminal region (TIR in the 5'-UTR and 5'-terminal fragments of the CDS) [18-20, 32-34]. This influence is often explained by the effect of the mRNA secondary structure on the efficiency of ribosome binding and translation initiation (Fig. 1) [19], which is believed to determine the overall protein synthesis rate [35]. It was shown that formation of stable hairpins in the mRNA 5'-terminal region decreases the efficiency of mRNA translation hundred times [36].

This hypothesis was confirmed by the discovery of mRNA secondary structures *in vivo* using the molecules selectively reacting with unpaired RNA bases, such as SHAPE [37] and DMS [38] probes. After the cells were treated with these compounds, their mRNA was sequenced in order to map modifications and to identify

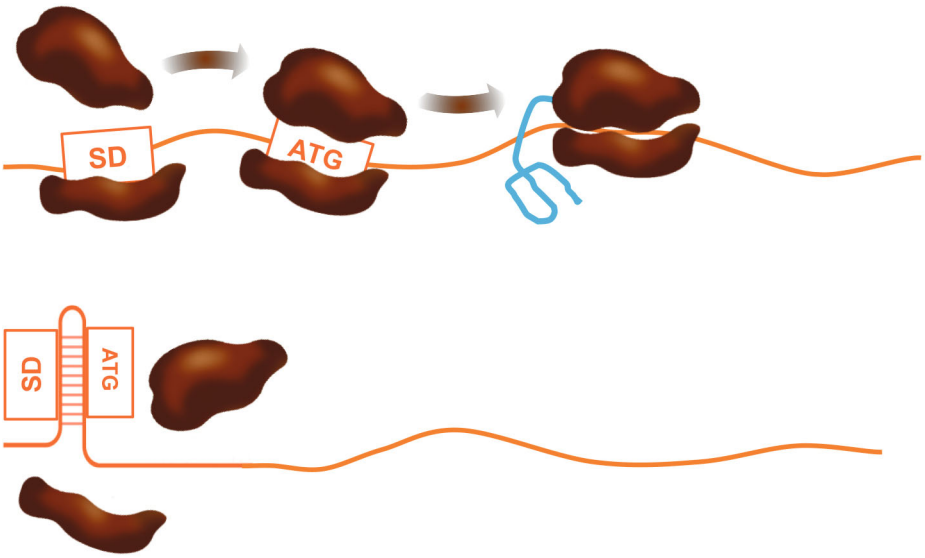

Fig. 1. The influence of the TIR secondary structure on the translation initiation efficiency. The ribosome easily binds to the SD sequence of mRNA lacking the secondary structure in the 5'-terminal region (top); the presence of the secondary structures in the region including the SD sequence hinders the binding of ribosomes (bottom).

unstructured and structured regions in the mRNA. One of the studies has shown that the efficiency of gene translation in *E. coli* cells was strongly determined by the absence in the TIR of the mRNA secondary structures that limited the availability of the SD sequence for the ribosomes [39, 40].

Translation elongation. Fast and slow codons. So far, no organisms with a full set of tRNAs with anticodons complementary to each of the 61 coding triplets have been found (e.g., *E. coli* has 39 tRNAs, Fig. 2) [41]. Translation of synonymous codons by the same tRNA occurs through the codon–anticodon interactions in which the first two bases of the codon form canonical Watson–Crick pairs (A/U, G/C) with the anticodon, whereas the third position allows the noncanonical pairing (G/U, A/I) due to wobbling. However, the affinity of the interaction between synonymous codons and the same tRNA is different. For instance, tRNAs of the 5'-GNN-3' type have a higher affinity toward the 5'-NNC-3' codons than toward the 5'-NNU-3' codons [42, 43]. Another important factor in the recognition of synonymous codons in the wobble hypothesis is modification of tRNAs. *E. coli* has very few codons that can be recognized by unmodified tRNAs [44]. Thus, the cmo^5U modification of the wobbling U base was detected in the tRNAs of the Ala, Leu, Pro, Ser, Thr, and Val codon families [45]. This modification allows the corresponding tRNA to recognize the codons with U, A, G, and C as the third base. Initially, it was believed that this modified base has an equal affinity toward any paired base; however, later studies have shown that its affinity to the codons ending with A or U is higher than for the codons ending with G or C [46], which affects the codon decoding rate.

Therefore, the kinetics of synonymous codon translation depends on numerous factors, such as: (i) the availability of the corresponding aminoacyl-tRNA, which depends on the gene copy number for this tRNA and the level of its expression [47, 48], (ii) the presence and the extent of tRNA modification [44, 49], (iii) the competition between different tRNAs for the codon reflecting their affinity and binding strength [50].

In general, a codon may be called slow if it slows down the process of elongation when decoded by the ribosome, independently of the mechanism of this phenomenon. Multiple attempts have been made to quantitatively assess the contribution of different codons to the rate and level of protein expression. The codon adaptation index (CAI) is the simplest parameter used to evaluate the representation of different codons in a gene of interest relatively to the codon composition of a set of standard genes with high expression levels [51]. This index is frequently used in the algorithms employed for the codon optimization for heterologous protein expression in *E. coli* [52]. However, it is far from perfect because of the following: (i) some mRNA regions display specific patterns of codon preference, often unrelated to the codon preference in the entire organism. Thus, the 5'-ends of CDSs of prokaryotic genes contain a region of ~15 codons that are often slow, which is important for proper initiation of elongation; (ii) the criteria for choosing the reference highly expressed genes are lacking as they depend on a large number of factors; (iii) in the case of heterologous expression, the expression machinery and specific features of codon preference can differ in the gene donor and host organisms.

To assess more accurately the rate of codon decoding, it was proposed to use the tRNA adaptation index (tAI) [53] that takes in account the number of gene copies for a tRNA specific for a given codon in the host cell (which is assumed to correlate with the tRNA content in the cell) and the efficiency of the codon binding with the anticodon according to the principles of Crick's wobbling. Although this index is more accurate, it does not take into account changes in the concentration of tRNAs (including aminoacyl-tRNAs and modified tRNAs) under different conditions. For instance, it was found that in bacteria, the extent of aminoacylation and modification of tRNAs recognizing synonymous codons can strongly fluctuate in response to amino acid starvation and depends on the cell division stage [54, 55]. The most recent and informative parameter that accounts not only for the total amount of tRNA, but also for the competition between the ribosomes for the tRNAs interacting with the same codon group, is the normalized translational efficiency (nTE) [56]. For a particular codon, nTE is the ratio of tAI to the translation frequency of this codon in an organism.

Although the elongation rate calculated based on tAI and nTE correlates well with the experimental data, further development of methods for the codon composition analysis should provide even more accurate predictions. The most important parameter for calculating the codon decoding rate is the concentration of mature tRNAs (aminoacyl-tRNAs and modified tRNAs) ready for the delivery of amino acids to the translation site. However, if the level of aminoacylation can be theoretically assessed from the gene copy number for a specific tRNA (as it is done for nTE and tAI), the extent of tRNA modification is difficult to evaluate. The mechanisms responsible for the tRNA modification in *E. coli* are poorly studied, but their importance in the regulation of mRNA translation is commonly recognized [49].

Translational ramp. Another specific feature of translation in *E. coli* cells is the preference for slow codons in the 5'-terminal region of the CDS termed the translational ramp [56, 57]. Analysis of the distribution of slow codons in *E. coli* genes revealed that the highly expressed genes contain ~10-15 relatively rare/slowly translated codons downstream of the start codon [56]. This finding has been confirmed in the experiments on ribosome profiling [32, 57, 58]. This approach is based

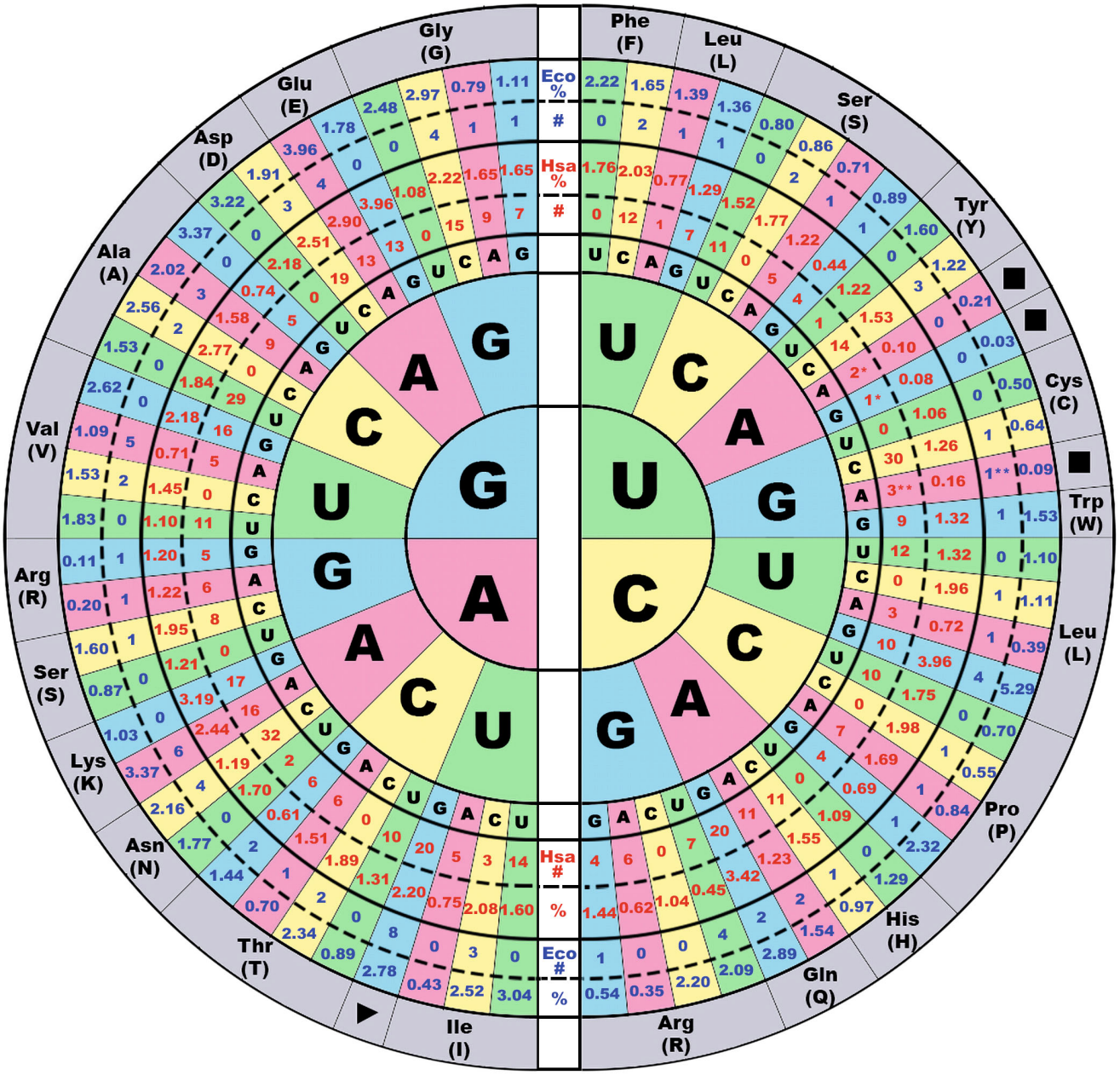

Fig. 2. Codon frequency (%) in all CDSs and the number of tRNA gene copies (#) in *E. coli* K12 (Eco, blue) and *Homo sapiens* (Hsa, red) (from http://gtrnadb.ucsc.edu). * Suppressor tRNAs for the stop codons. ** tRNA for selenocysteine and suppressor tRNAs for the TGA stop codon.

on the observation that the translating ribosome protects the mRNA region to which it is bound against the degradation by nucleases [12]. After suppressing translation with the translation elongation inhibitors, mRNA is treated with RNases, followed by sequencing of the mRNA fragments protected by the ribosomes. This allows to identify the position of the ribosome at the moment of translational arrest. These experiments showed an accumulation of ribosomes on the 5'-terminal region of the CDS, indicating a relatively low rate of translation in this region [32, 57, 58].

It is assumed that the slower ribosomal elongation speed leads to the uniform distribution of the ribosomes, which reduces ribosomal collisions and jamming during for the highly expressed genes with a relatively high ribosomal density (Fig. 3) [57, 58]. However, this finding can be explained otherwise: the evolutionary selection against the secondary structures at the mRNA 5'-end to facilitate translation initiation of the highly expressed genes was more important than the selection pressure toward the fast-translated codons [19]. In this case, the speed of ribosome movement after the translation initiation has to be balanced with the necessity for the absence of secondary structures at the mRNA 5'-end.

Synonymous codon ordering. The distribution of synonymous codons in the open reading frame is not random

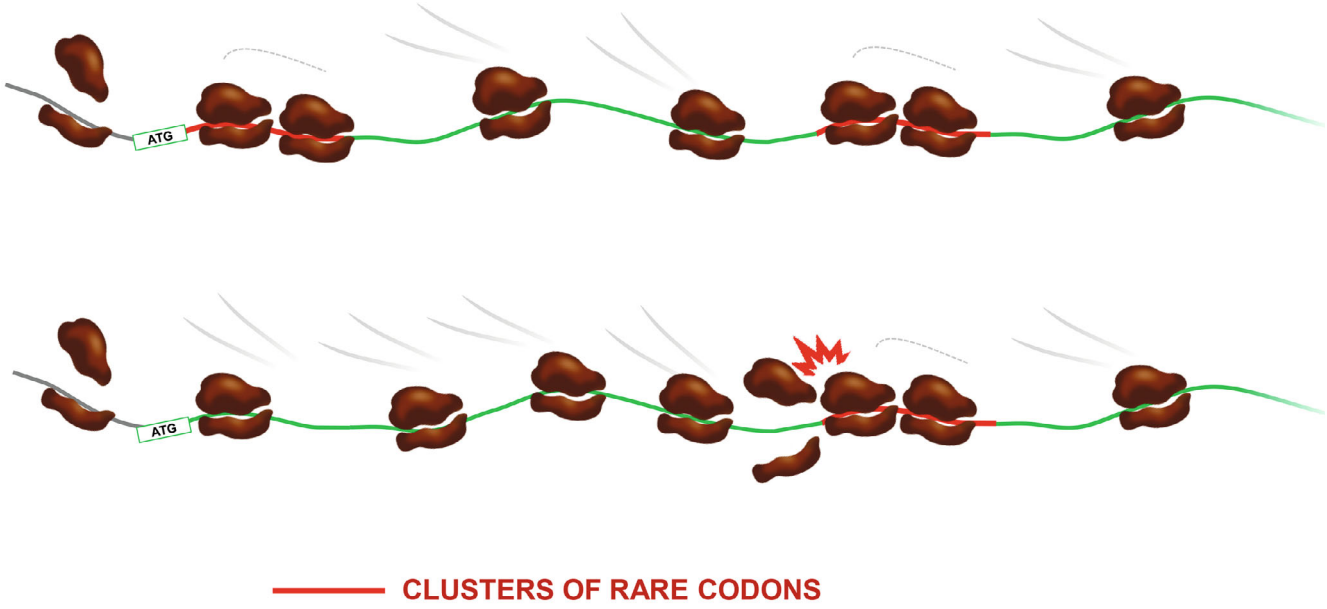

Fig. 3. Proposed effect of the translational ramp on translation; green, clusters of frequent codons in mRNA; red, clusters of rare codons. The proper distribution of rare codons in the region downstream of the start codon promotes a uniform distribution of ribosomes along the mRNA (top). When codon and/or host replacement affects the rate of the ribosome movement at the mRNA 5'-end, uneven distribution of the ribosomes might lead to the collision, jamming and premature termination of translation (below).

and follows a particular pattern [59]. It was shown that the arrangement of identical and some isoacceptor (recognized by the same tRNA) codons in the immediate proximity to each other usually favors translation.

This effect is believed to be associated with the fact that modification of the wobbling tRNA base affects not only the specificity, but also the affinity/efficiency of the tRNA molecule in the recognition of various codons. In theory, this can contribute to the development of the distribution patterns of synonymous codons (codon ordering) in bacteria under the pressure of evolutionary selection. Only identical pairs of codons and non-identical pairs in which the two codons are recognized with an equal (or close) high affinity by the same modified tRNA will favor translation and will be accumulated in bacterial genomes [59].

SD-like sequences are elements complementary or partially complementary to the anti-SD sequence of the 16S rRNA. It has been shown that the presence of SD-like sequences in bacterial CDSs can slow down the translation elongation and lead to a significant decrease in the protein production [60]. In most prokaryotes, including *E. coli*, the SD-like motifs are subjects of negative selection during the evolution [61, 62].

The AGG-AGG (Arg-Arg) pair is one of the most slowly translated codon pairs *in vivo*, presumably because of its significant affinity for the 16S rRNA fragment complementary to the SD sequence. This assumption has been confirmed by the finding that even an increase in the pool of the corresponding $tRNA^{ARG}_{AGG}$ due to the introduction of a multicopy plasmid containing gene (*argU*) for this tRNA failed to accelerate the translation of this pair [63, 64]. The placement of the AGG-AGG sequence into the 5'-terminal region of the CDS essentially decreased the protein expression level. The closer the introduced sequence was to the start codon, the stronger was its negative effect on protein expression [65].

Beside decreasing the protein expression level, a similar tandem of the codons (AGG-AGA) not only slowed down the translation, but also caused the premature termination with the production of a truncated protein [64]. Hence, these sequences should be avoided in the recombinant expression of genes in *E. coli*.

Ribosome collisions and jamming. In *E. coli* cells, the same mRNA molecule is simultaneously translated by several ribosomes with the formation of so-called polysome (Fig. 4). Polysomes can increase the translation efficiency by protecting the mRNA against degradation and increasing the time of its existence as a translation template [12]. Moreover, due to their helicase activity [66], ribosomes can destabilize the mRNA secondary structure, thus affecting the availability of the ribosome-binding site for the translation initiation [67, 68].

However, a high density of ribosomes on the mRNA in combination with the existence of rapidly translated regions and/or regions where translation is slowed down on the slow codons can lead to the ribosomal collision and jamming, which ultimately reduces the translation efficiency because of the translation slowing or even complete termination [69]. In particular, ribosomal collisions promote spontaneous dissociation of the jammed ribosomes or trigger the pathways leading to the dissociation of jammed ribosomes and mRNA degradation [70, 71].

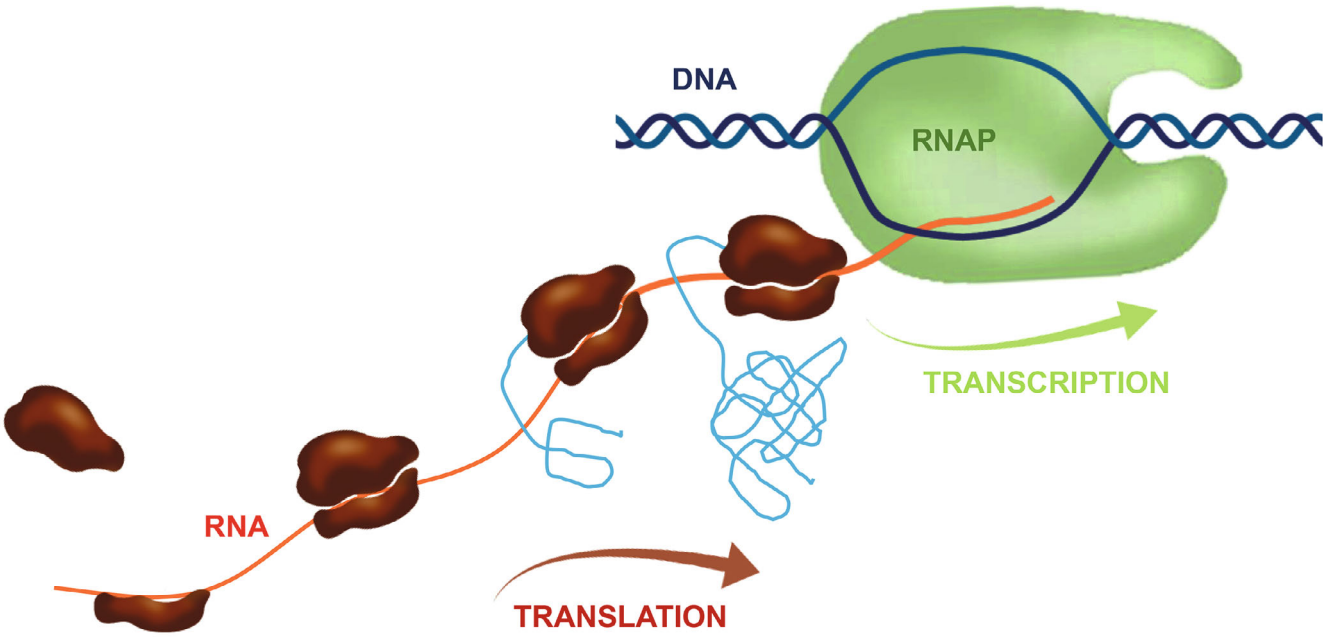

Fig. 4. Synchronous transcription/translation and formation of polysomes in bacteria.

Moreover, *E. coli* cells have a mechanism for releasing the jammed ribosomes that involves recognition of the stalled 50S subunits and proteolytic degradation of a partially formed polypeptide chain [72]. Therefore, bacteria possess the mechanisms for cotranslational degradation of mRNA and growing polypeptide that can also affect the translation efficiency.

The optimal distribution of ribosomes on the mRNA can be affected not only by the overall codon composition, but also by the translational ramp. Slow elongation at the early stages of translation can influence further uniform distribution of ribosomes along the entire mRNA molecule and thus prevent ribosomal collision and jamming [57].

Cotranslational folding of proteins. Slowing down elongation at certain translational stages can be critical for the proper folding of the growing polypeptide chain [73].

Protein folding *in vivo* starts during translation, when the leading peptide is released from the ribosomal tunnel. Variations in the local translation rate can promote local protein folding, allowing sequential structuring of domains of polypeptide chains emerging from the ribosome [74, 75]. A decrease in the translation rate increases the time required for the nascent polypeptide to fold correctly and to form the structural domains before the release of amino acid residues belonging to the other domains (Fig. 5). Alternatively, the acceleration of translation allows an entire domain to appear in a consistent manner, without formation of defective structures [76, 77].

Moreover, rare codons important for the protein folding were found not only in the unstructured interdomain regions, but also in the structured domains. Therefore, slowing down translation can be important for the folding of smaller structural subelements [78].

It has been shown experimentally that the replacement of rare codons by synonymous fast codons can lead to incorrect folding resulting in the protein aggregation (formation of inclusion bodies), degradation [79, 80], or emergence of proteins with altered functional properties. Thus, synonymous mutations can affect even the substrate specificity of enzymes [81], as well the phosphorylation profiles and activity of proteins [82].

It should be noted also that some structurally and/or functionally important protein fragments (e.g., enzyme active sites) can be encoded by frequent codons not because of the folding kinetics, but due to a more accurate translation of frequent codons that reduces the probability of error emergence in this fragment [83].

Termination of translation. Secondary structure of the mRNA 3'-end. It was found that the 3'-terminal regions of *E. coli* mRNAs have a decreased GC% content [33, 84], presumably due to the evolutionary selection against strong secondary structures that can affect normal termination of translation [33, 84-86]. This is evidenced by the correlation (although a poor one) between the reduced GC% content in the 3'-end of mRNA and increased protein expression in *E. coli* [33].There are also the differences in the codon usage preference in eukaryotes and prokaryotes: at the 3'-end of the gene CDSs preference for A/T-ending codons is more pronounced in bacteria than in eukaryotes [87].

Selection of stop codon. *E. coli* cells lack tRNAs capable of decoding the UAA, UAG, and UGA stop codons. Instead, UAG is decoded by the release factor 1 (RF1), UGA is decoded by RF2, and UAA is recognized by both RF1 and RF2. RF3 stimulates the termination of translation on all three stop codons [88, 89].

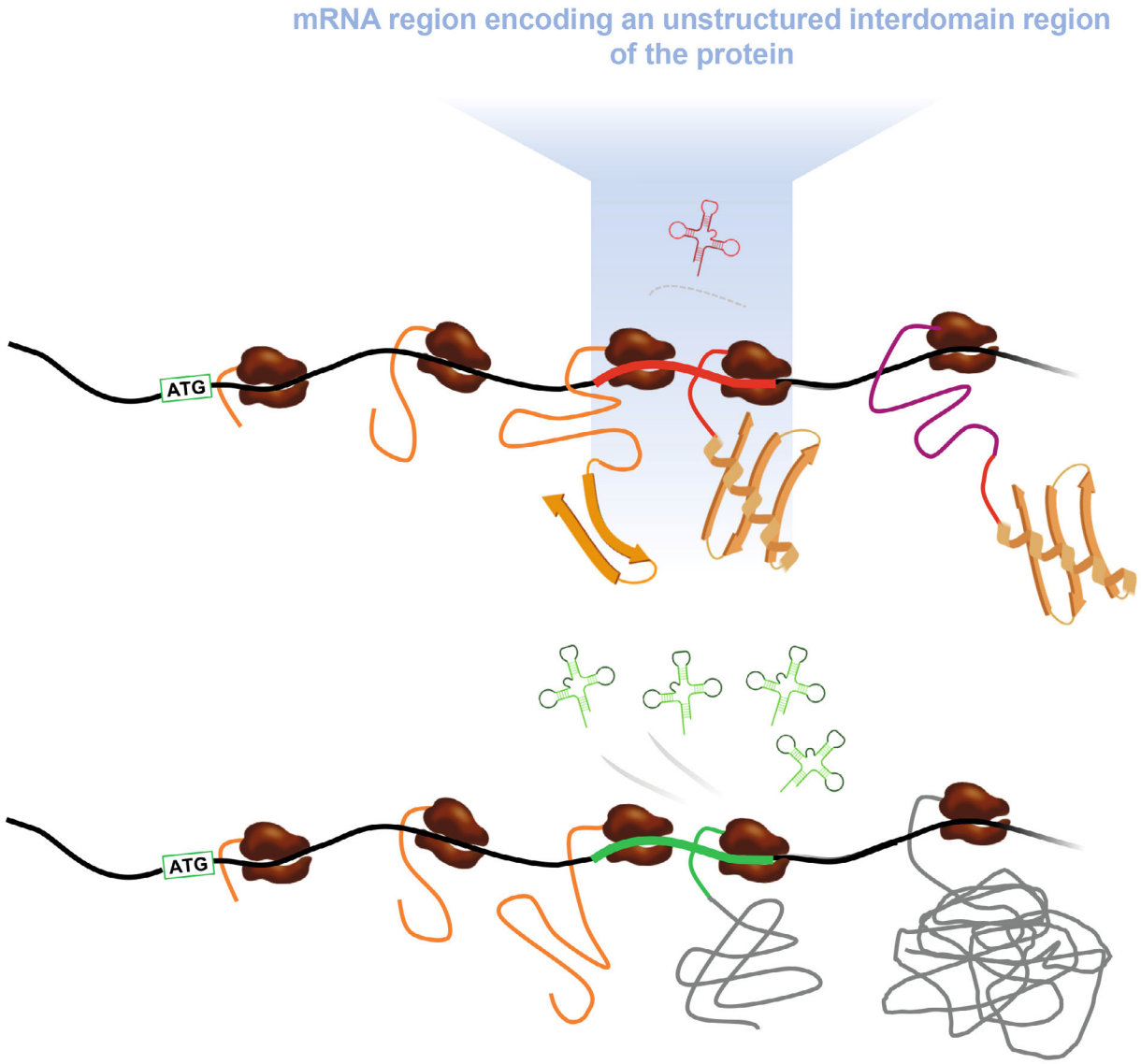

Fig. 5. Effect of changes in the translation rate of mRNA encoding an unstructured interdomain region on the folding of the entire protein molecule. When passing through the mRNA fragment encoding an unstructured interdomain region, the ribosomes can be slowed down to give time for the nascent polypeptide to fold properly (top); in the case of heterologous expression (host organism with different tRNA pool) or incorrect codon replacement in mRNA optimization, the ribosomes can pass this region faster, leading to the incorrect folding of the protein molecule (bottom).

The most frequently used stop codon in *E. coli* is UAA, whereas UAG is used least frequently [84, 90]. The base following the stop codon can be an important element of the translation termination signal. The termination efficiency markedly depends on the combination of the stop codon and the following fourth base: from 80% for UAAU, which is present along with UAAG in the majority of highly expressed genes, to 7% for the least efficient UGAC [91]. These data are typically explained by the contribution of the fourth base to the RF3 binding efficiency. Some research results also suggest that bases downstream of the stop codon (+4-+10) can also contribute to the termination efficiency by interacting with RF3 [92, 93], although no unambiguous evidence has been obtained yet to confirm this hypothesis.

It was suggested that the difference in the termination efficiency of different stop signals can be related to the rate and efficiency of the termination signal decoding [94], which, in turn, depends on the RF concentration for each stop codon, their binding affinity, and recycling rates [90]. Because the UAA codon is recognized by both RFs, the effect of the above parameters on the decoding efficiency for this codon is less pronounced. This is confirmed by the data that the efficiency of decoding strong stop signals UAAU and UAAG was significantly improved by the increase in the RF3 expression, while the efficiency of decoding weak signals was only modestly affected [95]. In the RF3-deficient cells, UAA demonstrated the highest efficiency and accuracy in comparison with UGA. The use of the UGA codon in these cells led to the increase in the number

of recoding events and ribosome accumulation near the stop codon [96].

Inefficient translation termination can induce ribosomal jamming, which might have a negative effect on the translation elongation, e.g., when several ribosomes accumulate on the translated mRNA before the stop codon [97]. Ribosomal jamming plays an especially noticeable role in the translation of short genes. Because of this, the most frequent stop codon in short, highly expressed genes in *E. coli* is UAA, instead of less efficient UGA [98]. Moreover, the dissociation of termination factors is important for the renovation of free ribosomal pool, which is essential for fast translation and promotion of protein expression [97].

THE USE OF CODON PREFERENCE

Improvement of expressing strain. As mentioned above, the rates of the codon decoding and, therefore, translation are often associated with the amount of tRNA decoding this codon. Several attempts have been made to increase the level of heterologous protein expression by introduction into the expressing strain of additional copies of the tRNA genes corresponding to the rare codons slowly translated in *E. coli*. Thus, *E. coli* Rosetta™ 2(DE3)pLysS strains contain the pRARE plasmid coding for tRNAs recognizing the AGG, AGA, AUA, CUA, CCC, and GGA codons. *E. coli* BL21-CodonPlus strain contains the pRIL plasmid coding for tRNAs specific for four codons rare in *E. coli* [99]. However, this approach solves only the problem associated with the slow ribosome movement and does not take into consideration the codon usage bias in different parts of the expressed gene.

Codon optimization. The codon preference for particular gene regions (codon usage bias) has formed in the course of evolution, and our major task now is to understand the principles for the efficient expression of proteins in heterologous systems. While achieving more comprehensive understanding of these principles and discovering the new ones, the scientists have gradually abandoned the concept that in order to achieve the high levels of protein expression, synthetic genes should contain as many frequent/rapidly translated codons as possible.

Standard constructs for increasing expression. The easiest approach to the expression optimization is the use of standard regulatory elements (promoters, SD sequences, TIRs, etc.) that could be reliably and reproducibly used in combination with the target genes to increase the expression levels of the recombinant proteins [6]. However, because of the lack of full understanding of all factors determining the efficiency of protein expression, the development and application of such constructs are only at the initial stage. Even in the well-known organisms, such as *E. coli*, apparently simple genetic constructs behave differently in different expression systems (i.e., using different expression constructs, strains, media, and cultivation conditions) [100].

Multiple experimental data have indicated that the efficiency of protein translation in *E. coli* is strongly determined by the mRNA 5'-UTR. As mentioned above, formation of secondary structures in this region is generally believed to be the main factor suppressing protein translation. These secondary structures can be formed not only by the mRNA 5'-UTR, but also with the involvement of 10-15 downstream codons in the CDS [6, 19, 20]. This noticeably hinders the development of standard constructs, because the SD sequences that work efficiently in the translation initiation for some CDSs sequences can be inactive with other CDSs [27]. Nevertheless, the attempts have been made to create the standard constructs by separating the two regions. To achieve stable gene expression, the authors of [101] used standardized modules, such as mRNA 5'-UTR and sequences coding for the N-terminal protein fragments that were cleaved off later to obtain the target protein [101]. In another study, the biscistronic modules have been successfully used for the separation of sequences responsible for the efficient translation initiation from the SD sequence and CDS [6, 18].

The creation of standard promoters is not an extremely challenging task, since it has been shown that the promoter strength does not vary strongly with different genes and remains at a predictable level [18].

Computer-assisted genes optimization. The most frequently used method of computer-assisted gene optimization is the adaptation of gene codon composition in accordance with the indices of codon usage in a selected expressing host [52]. For example, CAI is the most widespread method for analyzing codon usage that measures the deviation of a given CDS with respect to a set of highly expressed host reference genes (algorithms CAI calculator, CAIcal, CodonO, and CodonW).

Some laboratory-developed and commercial algorithms also take into consideration a number of additional parameters, such as the GC% content and the absence of specific motifs, such as the SD sequences, RNase E sites, or repeats with strong secondary structure within the CDS. Only several algorithms are directed to the minimization of secondary structures in the mRNA 5'-region, although this parameter is one of the key factors determining the efficiency of translation initiation and protein expression level [102].

Codon harmonization is another method of gene optimization. In this method, the codons in the expressed gene are replaced with the codons that have a similar translation rate in the heterologous expressing host [103], i.e., have a similar usage frequency [73, 82, 104]. This approach is frequently used for the optimization of protein folding in *E. coli* [105, 106].

Despite all the progress in this field, we still lack a comprehensive understanding of the codon composition influence on the expression and folding of heterologous proteins, which makes computer-assisted gene optimization very challenging. Many developed algorithms that account for multiple parameters (e.g., Eugene, DNA-Tailor) leave it for the user to establish the preferences for optimization, which is difficult because of the existing uncertainty about the importance of each parameter. This problem can be partly solved by using machine learning technologies which have been proven as good tools for the processing of large sets of related data. Various types of machine learning can be used for creating reliable algorithms to improve synthetic gene sequences. Thus, deep learning was used to create a novel codon optimization method in [107]. In this work, the training data including 4906 genes were selected from the DNA sequences of *E. coli* available from the NCBI database. Experimental verification has shown that this method was sufficiently efficient for increasing protein expression and commercially competitive. At present, machine learning methods are used mainly for the analysis of codon preferences in eukaryotes, e.g., Saccharomyces *cerevisiae* [108-110].

Such approaches are promising for improving prediction of protein expression level and folding. However, creation of machine learning algorithms requires the use of large datasets that have to be homogenous in order to avoid incorrect ratios between the parameters in the neuronal network. Another serious problem is that application of machine learning does not necessarily lead to a deeper understanding of biological processes. It is only a successful attempt to increase a given parameter, while deep neural network is a "black box" [107].

Therefore, enhancement of heterologous protein production using computer algorithms for codon optimization remains a method of trials and errors. Testing several variants increases the probability of success, but also increases the labor costs.

Library-based optimization methods. An alternative approach to the computer-assisted gene optimization and use of standard constructs is creation of libraries of synthetic plasmids with the randomized gene regions in need of optimization and the following screening of the resulting sequences for the expression levels. Synthetic libraries are widely used for the optimization of promoters [5, 6], SD sequences, TIRs [6, 18], and 5'-UTRs + CDSs [111, 112]. Along with the machine learning methods, randomized plasmid libraries are less dependent on our understanding of the mechanisms and specific features of codon preference because they allow to synchronously analyze multiple sequence variants. The creation of synthetic libraries and analysis of data obtained in such experiments also promote the development of new computer algorithms for gene optimization.

The main approach in the screening of libraries is the fusion of the CDS with the reporter protein. In this case, the level of protein expression and other parameters are evaluated based on the changes in the reporter signal intensity. Commonly used reporters are fluorescent proteins, such as mCherry, GFP, or superfolder GFP, which is a special form of GFP developed for expression of fusion proteins in *E. coli* [113]. Clones with the highest expression level of target protein are selected after plating on Petri dishes [114] or by cell sorting [19] based on the fluorescence signal intensity. Another type of reporter proteins are factors of antibiotic resistance. In this case, expressing cells are selected by screening the colonies grown on Petri dishes with a medium containing gradually increasing antibiotic concentrations. The clones with a higher level of the fused protein expression will demonstrate a high survival rate [111].

Some researchers believe that the presence of the reporter can distort the properties of the target protein, e.g., its solubility and expression level, which can lead to the false-positive or false-negative results. For this reason, recently developed TARSyn (tunable antibiotic resistance devices enabling bacterial synthetic evolution) system uses translational coupling devices sandwiched between the CDS and antibiotic selection marker using the BCD approach. This system has been demonstrated to ensure a highly productive selection of constructs with the optimized mRNA 5'-end for the expression of antibodies in *E. coli* [111].

However, the number of sequences that can be analyzed without highly efficient screening procedures is limited. On average, the complete degeneracy of 15 codons with no changes in the amino acid composition creates a library with ~2×10^7 variants that cannot be analyzed even by the most modern high-throughput methods. Therefore, the majority of studies using the library optimization methods (i) limit themselves to a lesser number of degenerated nucleotides or (ii) use the screening approaches that reduce the number of variants (for instance, TARSyn) [111]. Therefore, despite its obvious advantages, this approach seems to be most productive for the optimization of short gene regions, such as the SD sequences, promoters, TIRs, 5'-UTRs, etc., due to the used of limited-throughput assays insufficient for the analysis of large libraries and the absence of simple gene engineering methods for the randomization of long gene regions.

CONCLUSION

Development of gene optimization algorithms remains a challenging task that often limits the use of synthetic genes in biotechnology due to the problems of low protein yield or incorrect folding. Most gene optimization methods use obsolete parameters, such as CAI

or decrease in the number rare codon, without taking into account the results obtained in this field during the last years.

It has become obvious that using a single parameter for the optimization of an entire gene, without considering the codon preference bias, does not produce the desired result. The 5'- and 3'-ends of UTRs and CDS, domain boundaries, and other gene regions have different nucleotide preference patterns and can affect differently the level of protein expression and protein folding. Our current knowledge allows to some extent to create the constructs with the individually optimized gene regions, e.g., by using computer-assisted minimization of the mRNA 5'-end secondary structure (for increasing protein expression) [115] or by optimization of domain boundaries (for more efficient protein folding) [116, 117]. However, because of our incomplete understanding of the underlying processes, these methods still remain the trial-and-errors approaches. Therefore, creation of optimal synthetic constructs requires deeper comprehension of the codon preference for the individual gene regions and their combination.

The method of synthetic libraries is another widely used approach for the codon optimization, in which the optimization is achieved through the randomization of the regulatory and/or coding sequences followed by the analysis of the protein expression level based on the intensity of the reporter signal. The main problem of this approach is a rapid increase in the number of variants in the library with the increase in the randomized region size. Even analyzing the data after randomization of an individual region (for instance, 5'-UTR + CDS) is a nontrivial task and is limited to ~300,000 variants [18]. There are several ways to solve this problem. The first one is to use the regularities observed in previous experiments in order to limit the number of randomized variants and to decrease the library size. In theory, the principles of library construction obtained after several iterations should more adequately describe the principles of codon preference for a particular gene region, which will essentially simplify the analysis. Alternatively, the systems based on the use of reporter proteins can be used to limit the increase in the number of non-optimal variants. For example, the TARSyn system made it possible to select highly expressing clones based on their antibiotic resistance [111].

The development and use of new highly productive methods, such as the high-throughput sequencing, transcriptomics, and proteomics, also allow to obtain more significant and representative datasets that can be used for creating predictive theories and algorithms of gene optimization. The resulting large datasets can be analyzed using the machine learning algorithms [118]. Such approaches can be of help for more reliable identification of unknown functions and factors, as well as for the development of more elaborate algorithms for predicting protein expression levels.

In general, our understanding of the fundamental principles of codon preference and gene expression, as well as using this knowledge for solving the practical tasks, require more detailed studies. An important problem is the development of a model for evaluating the decoding rate of individual codons under normal conditions. The current metric parameters, such as tAI and nTE, ensure rather good prediction; however, they do not account for the content of aminoacyl-tRNAs and modified aminoacyl-tRNAs. Many organisms, including *E. coli*, can change the profiles of tRNA modification and aminoacylation under different growth conditions or under stress. From this point of view, the optimality of codons may be considered as a dynamic parameter (i.e., different fast or slow codons are required for an appropriate response to the changing conditions [119]) and, therefore, the experimental conditions should be normalized for the proper understanding of the corresponding processes. More accurate models for predicting the codon decoding rate can be useful for solving the problems of the ribosome movement speed, translational ramp, translational pauses, and translation-dependent folding.

Another, but no less important problem is the development of an accurate model of the gene segmentation accounting for the mutual effect of the codon composition of gene segments. As each gene segment has its own requirements for the codon optimization, it is important to find out how the codon preference of one segment affects the codon preference of another segment. When analyzing the effects of the CDS parameters on expression (e.g., the influence of the mRNA secondary structure on the expression level), it is important to minimize the interaction between the individual gene segments to assess the influence of each segment on the analyzed value. This approach will allow to reveal in more detail the codon preferences bias for the individual gene segments and/or their combinations.

Despite an impressive progress in protein expression, further improvement and development of new experimental and computational methods is essential for solving key problems in this field.

Contributions. Z.A.I. conceived the idea of manuscript. Z.A.I. and K.S.V. wrote the manuscript. K.A.G. supervised the project.

Ethics declarations. The authors declare no conflicts of interest. This article does not contain description of studies using humans or animal subjects performed by any of the authors.

REFERENCES

1. Huang, C. J., Lin, H., and Yang, X. (2012) Industrial production of recombinant therapeutics in *Escherichia coli*

and its recent advancements, *J. Industr. Microbiol. Biotechnol.*, **39**, 383-399, doi: 10.1007/s10295-011-1082-9.

2. Baeshen, M. N., Al-Hejin, A. M., Bora, R. S., Ahmed, M. M. M., Ramadan, H. A. I., Saini, K. S., and Redwan, E. M. (2015) Production of biopharmaceuticals in *E. coli*: current scenario and future perspectives, *J. Microbiol. Biotechnol., Korean Soc. Microbiol. Biotechnol.*, **25**, 953-962, doi: 10.4014/jmb.1412.12079.
3. Packiam, K. A. R., Ramanan, R. N., Ooi, C. W., Krishnaswamy, L., and Tey, B. T. (2020) Stepwise optimization of recombinant protein production in *Escherichia coli* utilizing computational and experimental approaches, *Appl. Microbiol. Biotechnol.*, **104**, 3253-3266, doi: 10.1007/s00253-020-10454-w.
4. Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Endy, D., and Church, G. M. (2013) Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*, *Proc. Natl. Acad. Sci. USA*, **110**, 14024-14029, doi: 10.1073/pnas.1301301110.
5. Alper, H., Fischer, C., Nevoigt, E., and Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering, *Proc. Natl. Acad. Sci. USA*, **102**, 12678-12683, doi: 10.1073/pnas.0504604102.
6. Mutalik, V. K., Guimaraes, J. C., Cambray, G., Lam, C., Christoffersen, M. J., Mai, Q. A., and Endy, D. (2013) Precise and reliable gene expression via standard transcription and translation initiation elements, *Nat. Methods*, **10**, 354-360, doi: 10.1038/nmeth.2404.
7. Conrad, T., Plumbom, I., Alcobendas, M., Vidal, R., and Sauer, S. (2020) Maximizing transcription of nucleic acids with efficient T7 promoters, *Commun. Biol.*, **3**, 439, doi: 10.1038/s42003-020-01167-x.
8. Komura, R., Aoki, W., Motone, K., Satomura, A., and Ueda, M. (2018) High-throughput evaluation of T7 promoter variants using biased randomization and DNA barcoding, *PLoS One*, **13**, e0196905, doi: 10.1371/journal.pone.0196905.
9. Villada, J. C., Duran, M. F., and Lee, P. K. H. (2020) Interplay between position-dependent codon usage bias and hydrogen bonding at the 5' end of ORFeomes, *mSystems*, **5**, e00613-20, doi: 10.1128/mSystems.00613-20.
10. Gorochowski, T. E., Chelysheva, I., Eriksen, M., Nair, P., Pedersen, S., and Ignatova, Z. (2019) Absolute quantification of translational regulation and burden using combined sequencing approaches, *Mol. Systems Biol.*, **15**, e8719, doi: 10.15252/msb.20188719.
11. Mittal, P., Brindle, J., Stephen, J., Plotkin, J. B., and Kudla, G. (2018) Codon usage influences fitness through RNA toxicity, *Proc. Natl. Acad. Sci. USA*, **115**, 8639-8644, doi: 10.1073/pnas.1810022115.
12. Dé Rique Braun, F., le Derout, J., and Ré Gnier, P. (1998) Ribosomes inhibit an RNase E cleavage which induces the decay of the rpsO mRNA of *Escherichia coli*, *EMBO J.*, **17**, 4790-4797, doi: 10.1093/emboj/17.16.4790.
13. Joel, J. (1993) *Control of Messenger RNA Stability. Part I: Prokaryotes. Part II: Eukaryotes*, Elsevier, pp. 495-517.
14. Kushner, S. R. (2004) mRNA decay in prokaryotes and eukaryotes: Different approaches to a similar problem, *IUBMB Life*, **56**, 585-594, doi: 10.1080/15216540400022441.
15. Emory, S. A., Bouvet, P., and Belasco, J. G. (1992) A 5'-terminal stem-loop structure can stabilize mRNA in *Escherichia coli*, *Genes Dev.*, **6**, 135-148, doi: 10.1101/gad.6.1.135.
16. Arnold, T. E., Yu, J., and Belasco, J. G. (1998) mRNA stabilization by the ompA 59 untranslated region: two protective elements hinder distinct pathways for mRNA degradation, *RNA*, **4**, 319-330.
17. Baker, K. E., and Mackie, G. A. (2003) Ectopic RNase E sites promote bypass of 5'-end-dependent mRNA decay in *Escherichia coli*, *Mol. Microbiol.*, **47**, 75-88, doi: 10.1046/j.1365-2958.2003.03292.x.
18. Cambray, G., Guimaraes, J. C., and Arkin, A. P. (2018) Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*, *Nat. Biotechnol.*, **36**, 1005-1015, doi: 10.1038/nbt.4238.
19. Goodman, D. B., Church, G. M., and Kosuri, S. (2013) Causes and effects of N-terminal codon bias in bacterial genes, *Science*, **342**, 475-479, doi: 10.1126/science.1241934.
20. Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009) Coding-sequence determinants of gene expression in *Escherichia coli*, *Science*, **324**, 255-258, doi: 10.1126/science.1170160.
21. Menendez-Gil, P., Caballero, C. J., Catalan-Moreno, A., Irurzun, N., Barrio-Hernandez, I., Caldelari, I., and Toledo-Arana, A. (2020) Differential evolution in 3'-UTRs leads to specific gene expression in *Staphylococcus*, *Nucleic Acids Res.*, **48**, 2544-2563, doi: 10.1093/nar/gkaa047.
22. Menendez-Gil, P., and Toledo-Arana, A. (2021) Bacterial 3'-UTRs: A Useful Resource in Post-transcriptional Regulation, *Frontiers Mol. Biosci.*, **7**, e617633, doi: 10.3389/fmolb.2020.617633.
23. Ruiz de los Mozos, I., Vergara-Irigaray, M., Segura, V., Villanueva, M., Bitarte, N., Saramago, M., and Toledo-Arana, A. (2013) Base pairing interaction between 5'- and 3'-UTRs controls icaR mRNA translation in *Staphylococcus aureus*, *PLoS Genet.*, **9**, e1004001, doi: 10.1371/journal.pgen.1004001.
24. Zhao, J. P., Zhu, H., Guo, X. P., and Sun, Y. C. (2018) AU-rich long 3' untranslated region regulates gene expression in bacteria, *Front. Microbiol.*, **9**, e3080, doi: 10.3389/fmicb.2018.03080.
25. McCarthy, J. E. G., and Gualerzi, C. (1990) Translational control of prokaryotic gene expression, *Trends Genet.*, **6**, 78-85, doi: 10.1016/0168-9525(90)90098-Q.
26. Laursen, B. S., Sørensen, H. P., Mortensen, K. K., and Sperling-Petersen, H. U. (2005) Initiation of protein synthesis in bacteria, *Microbiol. Mol. Biol. Rev.*, **69**, 101-123, doi: 10.1128/MMBR.69.1.101-123.2005.

27. Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009) Automated design of synthetic ribosome binding sites to control protein expression, *Nat. Biotechnol.*, **27**, 946-950, doi: 10.1038/nbt.1568.
28. Barrick, D., Villanueba, K., Childs, J., Kalil, R., Schneider, T. D., Lawrence, C. E., and Stormo, G. D. (1994) Quantitative analysis of ribosome binding sites in *E. coli*, *Nucleic Acids Res.*, **22**, 1287-1295, doi: 10.1093/nar/22.7.1287.
29. Nieuwkoop, T., Claassens, N. J., and van der Oost, J. (2019) Improved protein production and codon optimization analyses in *Escherichia coli* by bicistronic design, *Microb. Biotechnol.*, **12**, 173-179, doi: 10.1111/1751-7915.13332.
30. Schoner, B. E., Belagaje, R. M., and Schoner, R. G. (1986) Translation of a synthetic two-cistron mRNA in *Escherichia coli* (bovine growth hormone/human growth hormone/runaway replicon), *Biochemistry*, **83**, 8506-8510, doi: 10.1073/pnas.83.22.8506.
31. Makoff, A. J., and Smallwood, A. E. (1990) The use of two-cistron constructions in improving the expression of a heterologous gene in *E. coli*, *Nucleic Acids Res.*, **18**, 1711-1718, doi: 10.1093/nar/18.7.1711.
32. Boël, G., Letso, R., Neely, H., Price, W. N., Wong, K. H., Su, M., Luff, J., Valecha, M., Everett, J. K., Acton, T. B., Xiao, R., Montelione, G. T., Aalberts, D. P., and Hunt, J. F. (2016) Codon influence on protein expression in *E. coli* correlates with mRNA levels, *Nature*, **529**, 358-363, doi: 10.1038/nature16509.
33. Allert, M., Cox, J. C., and Hellinga, H. W. (2010) Multifactorial determinants of protein expression in prokaryotic open reading frames, *J. Mol. Biol.*, **402**, 905-918, doi: 10.1016/j.jmb.2010.08.010.
34. Del Campo, C., Bartholomäus, A., Fedyunin, I., and Ignatova, Z. (2015) Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function, *PLoS Genet.*, **11**, e1005613, doi: 10.1371/journal.pgen.1005613.
35. Kozak, M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes, *Gene*, **361**, 13-37, doi: 10.1016/j.gene.2005.06.037.
36. Borujeni, A. E., Cetnar, D., Farasat, I., Smith, A., Lundgren, N., and Salis, H. M. (2017) Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences, *Nucleic Acids Res.*, **45**, 5437-5448, doi: 10.1093/nar/gkx061.
37. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E., and Weeks, K. M. (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP), *Nat. Methods*, **11**, 959-965, doi: 10.1038/nmeth.3029.
38. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J. S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*, *Nature*, **505**, 701-705, doi: 10.1038/nature12894.
39. Mustoe, A. M., Busan, S., Rice, G. M., Hajdin, C. E., Peterson, B. K., Ruda, V. M., Kubica, N., Nutiu, R., Baryza, J. L., and Weeks, K. M. (2018) Pervasive regulatory functions of mRNA structure revealed by high-resolution SHAPE probing, *Cell*, **173**, 181-195.e18, doi: 10.1016/j.cell.2018.02.034.
40. Kelsic, E. D., Chung, H., Cohen, N., Park, J., Wang, H. H., and Kishony, R. (2016) RNA structural determinants of optimal codons revealed by MAGE-Seq, *Cell Systems*, **3**, 563-571.e6, doi: 10.1016/j.cels.2016.11.004.
41. Gouy, M., and Grantham, R. (1980) Polypeptide elongation and tRNA cycling in *Escherichia coli*: a dynamic approach, *FEBS Lett.*, **115**, 151-155, doi: 10.1016/0014-5793(80)81155-0.
42. Crick, F. H. C. (1966) Codon-anticodon pairing: the wobble hypothesis, *J. Mol. Biol.*, **19**, 548-555, doi: 10.1016/S0022-2836(66)80022-0.
43. Söll, D., Jones, D. S., Ohtsuka, E., Faulkner, R. D., Lohrmann, R., Hayatsu, H., and Khorana, H. G. (1966) Specificity of sRNA for recognition of codons as studied by the ribosomal binding technique, *J. Mol. Biol.*, **19**, 556-573, doi: 10.1016/S0022-2836(66)80023-2.
44. Agris, P. F., Vendeix, F. A. P., and Graham, W. D. (2007) tRNA's wobble decoding of the genome: 40 years of modification, *J. Mol. Biol.*, **366**, 1-13, doi: 10.1016/j.jmb.2006.11.046.
45. Kothe, U., and Rodnina, M. V. (2007) Codon reading by tRNAAla with modified uridine in the wobble position, *Mol. Cell*, **25**, 167-174, doi: 10.1016/j.molcel.2006.11.014.
46. Ran, W., and Higgs, P. G. (2010) The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria, *Mol. Biol. Evol.*, **27**, 2129-2140, doi: 10.1093/molbev/msq102.
47. Dykeman, E. C. (2020) A stochastic model for simulating ribosome kinetics *in vivo*, *PLoS Computat. Biol.*, **16**, e1007618, doi: 10.1371/journal.pcbi.1007618.
48. Vieira, J. P., Racle, J., and Hatzimanikatis, V. (2016) Analysis of translation elongation dynamics in the context of an *Escherichia coli* cell, *Biophys. J.*, **110**, 2120-2131, doi: 10.1016/j.bpj.2016.04.004.
49. De Crécy-Lagard, V., and Jaroch, M. (2021) Functions of bacterial tRNA modifications: from ubiquity to diversity, *Trends Microbiol.*, **29**, 41-53, doi: 10.1016/j.tim.2020.06.010.
50. Gromadski, K. B., Daviter, T., and Rodnina, M. V. (2006) A uniform response to mismatches in codon-anticodon complexes ensures ribosomal fidelity, *Mol. Cell*, **21**, 369-377, doi: 10.1016/j.molcel.2005.12.018.
51. Sharp, P. M., and Li, W. H. (1987) The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Res.*, **15**, 1281-1295, doi: 10.1093/nar/15.3.1281.
52. Parret, A. H., Besir, H., and Meijers, R. (2016) Critical reflections on synthetic gene design for recombinant protein expression, *Curr. Opin. Structur. Biol.*, **38**, 155-162, doi: 10.1016/j.sbi.2016.07.004.
53. Dos Reis, M., Savva, R., and Wernisch, L. (2004) Solving the riddle of codon usage preferences: A test for trans-

lational selection, *Nucleic Acids Res.*, **32**, 5036-5044, doi: 10.1093/nar/gkh834.
54. Elf, J. (2003) Selective Charging of tRNA isoacceptors explains patterns of codon usage, *Science*, **300**, 1718-1722, doi: 10.1126/science.1083811.
55. Dittmar, K. A., Sørensen, M. A., Elf, J., Ehrenberg, M., and Pan, T. (2005) Selective charging of tRNA isoacceptors induced by amino-acid starvation, *EMBO Rep.*, **6**, 151-157, doi: 10.1038/sj.embor.7400341.
56. Pechmann, S., and Frydman, J. (2013) Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding, *Nat. Struct. Mol. Biol.*, **20**, 237-243, doi: 10.1038/nsmb.2466.
57. Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., and Pilpel, Y. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation, *Cell*, **141**, 344-354, doi: 10.1016/j.cell.2010.03.031.
58. Tuller, T., and Zur, H. (2015) Multiple roles of the coding sequence 5'-end in gene expression regulation, *Nucleic Acids Res.*, **43**, 13-28, doi: 10.1093/nar/gku1313.
59. Shao, Z. Q., Zhang, Y. M., Feng, X. Y., Wang, B., and Chen, J. Q. (2012) Synonymous codon ordering: A subtle but prevalent strategy of bacteria to improve translational efficiency, *PLoS One*, **7**, e33547, doi: 10.1371/journal.pone.0033547.
60. Li, G. W., Oh, E., and Weissman, J. S. (2012) The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria, *Nature*, **484**, 538-541, doi: 10.1038/nature10965.
61. Diwan, G. D., and Agashe, D. (2016) The frequency of internal Shine–Dalgarno-like motifs in prokaryotes, *Genome Biol. Evol.*, **8**, 1722-1733, doi: 10.1093/gbe/evw107.
62. Hockenberry, A. J., Jewett, M. C., Amaral, L. A. N., and Wilke, C. O. (2018) Within-gene Shine–Dalgarno sequences are not selected for function, *Mol. Biol. Evol.*, **35**, 2487-2498, doi: 10.1093/molbev/msy150.
63. Chevance, F. F. V., le Guyon, S., and Hughes, K. T. (2014) The Effects of codon context on *in vivo* translation speed, *PLoS Genet.*, **10**, e1004392, doi: 10.1371/journal.pgen.1004392.
64. Correddu, D., Montaño López, J. de J., Angermayr, S. A., Middleditch, M. J., Payne, L. S., and Leung, I. K. H. (2020) Effect of consecutive rare codons on the recombinant production of human proteins in *Escherichia coli*, *IUBMB Life*, **72**, 266-274, doi: 10.1002/iub.2162.
65. Osterman, I. A., Chervontseva, Z. S., Evfratov, S. A., Sorokina, A. v., Rodin, V. A., Rubtsova, M. P., and Sergiev, P. V. (2020) Translation at first sight: the influence of leading codons, *Nucleic Acids Res.*, **48**, 6931-6942, doi: 10.1093/nar/gkaa430.
66. Takyar, S., Hickerson, R. P., and Noller, H. F. (2005) mRNA helicase activity of the ribosome, *Cell*, **120**, 49-58, doi: 10.1016/j.cell.2004.11.042.
67. Andreeva, I., Belardinelli, R., and Rodnina, M. V. (2018) Translation initiation in bacterial polysomes through ribosome loading on a standby site on a highly translated mRNA, *Proc. Natl. Acad. Sci. USA*, **115**, 4411-4416, doi: 10.1073/pnas.1718029115.
68. Burkhardt, D. H., Rouskin, S., Zhang, Y., Li, G.-W., Weissman, J. S., and Gross, C. A. (2017) Operon mRNAs are organized into ORF-centric structures that predict translation efficiency, *eLife*, **6**, e22037, doi: 10.7554/eLife.22037.
69. Keiler, K. C. (2015) Mechanisms of ribosome rescue in bacteria, *Nat. Rev. Microbiol.*, **13**, 285-297, doi: 10.1038/nrmicro3438.
70. Janssen, B. D., and Hayes, C. S. (2012) The tmRNA ribosome-rescue system, *Adv. Protein Chem. Struct. Biol.*, **86**, 151-191, doi: 10.1016/B978-0-12-386497-0.00005-0.
71. Moore, S. D., and Sauer, R. T. (2007) The tmRNA system for translational surveillance and ribosome rescue, *Ann. Rev. Biochem.*, **76**, 101-124, doi: 10.1146/annurev.biochem.75.103004.142733.
72. Lytvynenko, I., Paternoga, H., Thrun, A., Balke, A., Müller, T. A., Chiang, C. H., Nagler, K., Tsaprailis, G., Anders, S., Bischofs, I., Maupin-Furlow, J. A., Spahn, C. M. T., and Joazeiro, C. A. P. (2019) Alanine tails signal proteolysis in bacterial ribosome-associated quality control, *Cell*, **178**, 76-90.e22, doi: 10.1016/j.cell.2019.05.002.
73. Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M. V., and Komar, A. A. (2016) Synonymous codons direct cotranslational folding toward different protein conformations, *Mol. Cell*, **61**, 341-351, doi: 10.1016/j.molcel.2016.01.008.
74. Waudby, C. A., Launay, H., Cabrita, L. D., and Christodoulou, J. (2013) Protein folding on the ribosome studied using NMR spectroscopy, *Progr. Nucl. Magn. Reson. Spectrosc.*, **74**, 57-75, doi: 10.1016/j.pnmrs.2013.07.003.
75. Gloge, F., Becker, A. H., Kramer, G., and Bukau, B. (2014) Co-translational mechanisms of protein maturation, *Curr. Opin. Struct. Biol.*, **24**, 24-33, doi: 10.1016/j.sbi.2013.11.004.
76. Jacobson, G. N., and Clark, P. L. (2016) Quality over quantity: Optimizing co-translational protein folding with non-"optimal" synonymous codons, *Curr. Opin. Struct. Biol.*, **38**, 102-110, doi: 10.1016/j.sbi.2016.06.002.
77. Sander, I. M., Chaney, J. L., and Clark, P. L. (2014) Expanding anfinsen's principle: contributions of synonymous codon selection to rational protein design, *J. Am. Chem. Soc.*, **136**, 858-861, doi: 10.1021/ja411302m.
78. Chaney, J. L., Steele, A., Carmichael, R., Rodriguez, A., Specht, A. T., Ngo, K., and Clark, P. L. (2017) Widespread position-specific conservation of synonymous rare codons within coding sequences, *PLoS Comput. Biol.*, **13**, e1005531, doi: 10.1371/journal.pcbi.1005531.
79. Spencer, P. S., Siller, E., Anderson, J. F., and Barral, J. M. (2012) Silent substitutions predictably alter translation elongation rates and protein folding efficiencies, *J. Mol. Biol.*, **422**, 328-335, doi: 10.1016/j.jmb.2012.06.010.
80. Zhang, G., Hubalewska, M., and Ignatova, Z. (2009) Transient ribosomal attenuation coordinates protein

synthesis and co-translational folding, *Nat. Struct. Mol. Biol.*, **16**, 274-280, doi: 10.1038/nsmb.1554.
81. Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., and Gottesman, M. M. (2007) A "Silent" polymorphism in the *MDR1* gene changes substrate specificity, *Science*, **315**, 525-528, doi: 10.1126/science.1135308.
82. Zhou, M., Guo, J., Cha, J., Chae, M., Chen, S., Barral, J. M., and Liu, Y. (2013) Non-optimal codon usage affects expression, structure and function of clock protein FRQ, *Nature*, **494**, 111-115, doi: 10.1038/nature11833.
83. Drummond, D. A., and Wilke, C. O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution, *Cell*, **134**, 341-352, doi: 10.1016/j.cell.2008.05.042.
84. Eyre-Walker, A. (1996) The close proximity of *Escherichia coli* genes: consequences for stop codon and synonymous codon use, *J. Mol. Evol.*, **42**, 73-78, doi: 10.1007/BF02198830.
85. Katz, L., and Burge, C. B. (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes, *Genome Res.*, **13**, 2042-2051, doi: 10.1101/gr.1257503.
86. Rocha, E. P. C., Danchin, A., and Viari, A. (1999) Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis, *Nucleic Acids Res.*, **27**, 3567-3576, doi: 10.1093/nar/27.17.3567.
87. Zahdeh, F., and Carmel, L. (2019) Nucleotide composition affects codon usage toward the 3'-end, *PLoS One*, **14**, e0225633, doi: 10.1371/journal.pone.0225633.
88. Capecchi, M. R. (1967) Polypeptide chain termination *in vitro*: isolation of a release factor, *Proc. Natl. Acad. Sci. USA*, **58**, 1144-1151, doi: 10.1073/pnas.58.3.1144.
89. Scolnick, E. M., and Caskey, C. T. (1969) Peptide chain termination, V. The role of release factors in mRNA terminator codon recognition, *Proc. Natl. Acad. Sci. USA*, **64**, 1235-1241, doi: 10.1073/pnas.64.4.1235.
90. Korkmaz, G., Holm, M., Wiens, T., and Sanyal, S. (2014) Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance, *J. Biol. Chem.*, **289**, 30334-30342, doi: 10.1074/jbc.M114.606632.
91. Poole, E. S., Brown, C. M., and Tate, W. P. (1995) The identity of the base following the stop codon determines the efficiency of *in vivo* translational termination in *Escherichia coli*, *EMBO J.*, **14**, 151-158, doi: 10.1002/j.1460-2075.1995.tb06985.x.
92. Tate, W. P., and Mannering, S. A. (1996) Three, four or more: the translational stop signal at length, *Mol. Microbiol.*, **21**, 213-219, doi: 10.1046/j.1365-2958.1996.6391352.x.
93. Namy, O., Hatin, I., and Rousset, J.-P. (2001) Impact of the six nucleotides downstream of the stop codon on translation termination, *EMBO Rep.*, **2**, 2001, doi: 10.1093/embo-reports/kve176.
94. Sharp, P. M., and Bulmer, M. (1988) Selective differences among translation termination codons, *Gene*, **63**, 141-145, doi: 10.1016/0378-1119(88)90553-7.
95. Crawford, D. J. G., Ito, K., Nakamura, Y., and Tate, W. P. (1999) Indirect regulation of translational termination efficiency at highly expressed genes and recoding sites by the factor recycling function of *Escherichia coli* release factor RF3, *EMBO J.*, **18**, 727-732, doi: 10.1093/emboj/18.3.727.
96. Baggett, N. E., Zhang, Y., and Gross, C. A. (2017) Global analysis of translation termination in *E. coli*, *PLoS Genet.*, **13**, e1006676, doi: 10.1371/journal.pgen.1006676.
97. Pavlov, M. Yu., Freistroffer, D. V., Dincbas, V., MacDougall, J., Buckingham, R. H., and Ehrenberg, M. (1998) A direct estimation of the context effect on the efficiency of termination, *J. Mol. Biol.*, **284**, 579-590, doi: 10.1006/jmbi.1998.2220.
98. Jin, H. (2002) Cis control of gene expression in *E. coli* by ribosome queuing at an inefficient translational stop signal, *EMBO J.*, **21**, 4357-4367, doi: 10.1093/emboj/cdf424.
99. Gustafsson, C., Govindarajan, S., and Minshull, J. (2004) Codon bias and heterologous protein expression, *Trends Biotechnol.*, **22**, 346-353, doi: 10.1016/j.tibtech.2004.04.006.
100. Kittleson, J. T., Wu, G. C., and Anderson, J. C. (2012) Successes and failures in modular genetic engineering, *Curr. Opin. Chem. Biol.*, **16**, 329-336, doi: 10.1016/j.cbpa.2012.06.009.
101. Ki, M. R., and Pack, S. P. (2020) Fusion tags to enhance heterologous protein expression, *Appl. Microbiol. Biotechnol.*, **104**, 2411-2425, doi: 10.1007/s00253-020-10402-8.
102. Gould, N., Hendy, O., and Papamichail, D. (2014) Computational tools and algorithms for designing customized synthetic genes, *Front. Bioeng. Biotechnol.*, **2**, e00041, doi: 10.3389/fbioe.2014.00041.
103. Tian, J., Yan, Y., Yue, Q., Liu, X., Chu, X., Wu, N., and Fan, Y. (2017) Predicting synonymous codon usage and optimizing the heterologous gene for expression in *E. coli*, *Sci. Rep.*, **7**, e9926, doi: 10.1038/s41598-017-10546-0.
104. Rodriguez, A., Wright, G., Emrich, S., and Clark, P. L. (2018) %MinMax: A versatile tool for calculating and comparing synonymous codon usage and its impact on protein folding, *Protein Sci.*, **27**, 356-362, doi: 10.1002/pro.3336.
105. Angov, E. (2011) Codon usage: Nature's roadmap to expression and folding of proteins, *Biotechnol. J.*, **6**, 650-659, doi: 10.1002/biot.201000332.
106. Hillier, C. J., Ware, L. A., Barbosa, A., Angov, E., Lyon, J. A., Heppner, D. G., and Lanar, D. E. (2005) Process development and analysis of liver-stage antigen 1, a pre-erythrocyte-stage protein-based vaccine for *Plasmodium falciparum*, *Infect. Immun.*, **73**, 2109-2115, doi: 10.1128/IAI.73.4.2109-2115.2005.
107. Fu, H., Liang, Y., Zhong, X., Pan, Z. L., Huang, L., Zhang, H. L., Xu, Y., Zhou, W., and Liu, Z. (2020) Codon optimization with deep learning to enhance protein expression, *Sci. Rep.*, **10**, 17617, doi: 10.1038/s41598-020-74091-z.

108. Cuperus, J. T., Groves, B., Kuchina, A., Rosenberg, A. B., Jojic, N., Fields, S., and Seelig, G. (2017) Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences, *Genome Res.*, **27**, 2015-2024, doi: 10.1101/gr.224964.117.
109. Decoene, T., Peters, G., de Maeseneire, S. L., and de Mey, M. (2018) Toward PredicTable 5'UTRs in *Saccharomyces cerevisiae*: development of a yUTR calculator, *ACS Synthet. Biol.*, **7**, 622-634, doi: 10.1021/acssynbio.7b00366.
110. de Jongh, R. P. H., van Dijk, A. D. J., Julsing, M. K., Schaap, P. J., and de Ridder, D. (2020) Designing eukaryotic gene expression regulation using machine learning, *Trends Biotechnol.*, **38**, 191-201, doi: 10.1016/j.tibtech.2019.07.007.
111. Rennig, M., Martinez, V., Mirzadeh, K., Dunas, F., Röjsäter, B., Daley, D. O., and Nørholm, M. H. H. (2018) TARSyn: tunable antibiotic resistance devices enabling bacterial synthetic evolution and protein production, *ACS Synthet. Biol.*, **7**, 432-442, doi: 10.1021/acssynbio.7b00200.
112. Mirzadeh, K., Martínez, V., Toddo, S., Guntur, S., Herrgård, M. J., Elofsson, A., and Daley, D. O. (2015) Enhanced protein production in *Escherichia coli* by optimization of cloning scars at the vector-coding sequence junction, *ACS Synthet. Biol.*, **4**, 959-965, doi: 10.1021/acssynbio.5b00033.
113. Pédelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C., and Waldo, G. S. (2006) Engineering and characterization of a superfolder green fluorescent protein, *Nat. Biotechnol.*, **24**, 79-88, doi: 10.1038/nbt1172.
114. Mirzadeh, K., Shilling, P. J., Elfageih, R., Cumming, A. J., Cui, H. L., Rennig, M., and Daley, D. O. (2020) Increased production of periplasmic proteins in *Escherichia coli* by directed evolution of the translation initiation region, *Microb. Cell Factor.*, **19**, e85, doi: 10.1186/s12934-020-01339-8.
115. Care, S., Bignon, C., Pelissier, M. C., Blanc, E., Canard, B., and Coutard, B. (2008) The translation of recombinant proteins in *E. coli* can be improved by *in silico* generating and screening random libraries of a −70/+96 mRNA region with respect to the translation initiation codon, *Nucleic Acids Res.*, **36**, e6, doi: 10.1093/nar/gkm1097.
116. Hess, A.-K., Saffert, P., Liebeton, K., and Ignatova, Z. (2015) Mathematisch-Naturwissenschaftliche Fakultät Optimization of translation profiles enhances protein expression and solubility, *PLOS One*, **10**, e127039, doi: 10.1371/journal.pone.0127039.
117. Zhong, C., Wei, P., and Zhang, Y. H. P. (2017) Enhancing functional expression of codon-optimized heterologous enzymes in *Escherichia coli* BL21(DE3) by selective introduction of synonymous rare codons, *Biotechnol. Bioeng.*, **114**, 1054-1064, doi: 10.1002/bit.26238.
118. Zrimec, J., Buric, F., Kokina, M., Garcia, V., and Zelezniak, A. (2021) Learning the regulatory code of gene expression, *Front. Mol. Biosci.*, **8**, e673363, doi: 10.3389/fmolb.2021.673363.
119. Hanson, G., and Coller, J. (2018) Codon optimality, bias and usage in translation and mRNA decay, *Nat. Rev. Mol. Cell Biol.*, **19**, 20-30, doi: 10.1038/nrm.2017.91.