PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Evolutionary analysis of polyproline motifs in *Escherichia coli* reveals their regulatory role in translation

Fei Qi[1], Magdalena Motz[2,3], Kirsten Jung[2,3], Jürgen Lassak[2,3], Dmitrij Frishman[1,4]*

1 Department of Bioinformatics, Wissenschaftzentrum Weihenstephan, Technische Universität München, Freising, Germany, **2** Center for Integrated Protein Science Munich, Ludwig-Maximilians-Universität München, Munich, Germany, **3** Department of Biology I, Microbiology, Ludwig-Maximilians-Universität München, Martinsried, Germany, **4** St Petersburg State Polytechnic University, St Petersburg, Russia

* d.frishman@wzw.tum.de

Abstract

Translation of consecutive prolines causes ribosome stalling, which is alleviated but cannot be fully compensated by the elongation factor P. However, the presence of polyproline motifs in about one third of the *E. coli* proteins underlines their potential functional importance, which remains largely unexplored. We conducted an evolutionary analysis of polyproline motifs in the proteomes of 43 *E. coli* strains and found evidence of evolutionary selection against translational stalling, which is especially pronounced in proteins with high translational efficiency. Against the overall trend of polyproline motif loss in evolution, we observed their enrichment in the vicinity of translational start sites, in the inter-domain regions of multi-domain proteins, and downstream of transmembrane helices. Our analysis demonstrates that the time gain caused by ribosome pausing at polyproline motifs might be advantageous in protein regions bracketing domains and transmembrane helices. Polyproline motifs might therefore be crucial for co-translational folding and membrane insertion.

Author summary

Polyproline motifs induce ribosome stalling during translation, but the functional significance of this effect remains unclear. Our evolutionary analysis of polyproline motifs reveals that they are disfavored in *E. coli* proteomes as a consequence of the reduced translation efficiency, supporting the conjecture that translation efficiency-based evolutionary pressure shapes protein sequences. Enrichment of polyproline motifs in the protein regions bracketing structural domains and transmembrane helices indicates their regulatory role in co-translational protein folding and transmembrane helix insertion. Polyproline motifs could thus serve as protein-level *cis*-acting elements, which directly regulate the rate of translation elongation.

Citation: Qi F, Motz M, Jung K, Lassak J, Frishman D (2018) Evolutionary analysis of polyproline motifs in *Escherichia coli* reveals their regulatory role in translation. PLoS Comput Biol 14(2): e1005987. https://doi.org/10.1371/journal.pcbi.1005987

Editor: Charlotte M. Deane, University of Oxford, UNITED KINGDOM

Received: September 28, 2017

Accepted: January 17, 2018

Published: February 1, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the DFG: Center for Integrated Protein Science Munich (CIPSM) (http://www.dfg.de/; Exc114/2 and TRR174/P09 to KJ, and LA 3658/1-1 to JL) and the China Scholarship Council (http://www.csc.edu.cn/; 201206740006 to FQ). This work was supported by the German Research Foundation (DFG) and the Technical University of Munich (TUM) in the framework of the Open Access Publishing Program. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Ribosomes facilitate the synthesis of proteins by translating the nucleotide sequence from an mRNA template. The speed of mRNA translation significantly varies and strongly depends on the amino acids to be incorporated into the growing polypeptide chain [1]. Especially slow is the incorporation of proline [2–4]. The pyrrolidine ring gives proline an exceptional conformational rigidity compared to all other amino acids and makes it not only a poor A-site peptidyl acceptor [2], but also a poor P-site peptidyl donor [3,4]. Translation of two and more consecutive prolines dramatically impairs the peptidyl transfer reaction and eventually causes ribosomes to stall [3,5–9]. Although basically all diproline comprising motifs cause translational stalling [5,10], the arrest strength is influenced by physical and chemical properties of the adjacent amino acids that affect the conformation of the nascent polypeptide chain. Based on proteomic approaches combined with systematic *in vivo* and *in vitro* analyses, a hierarchy of arrest peptides was described [5,9–11]. Thereby triplets such as PPP, D/PP/D, PPW, APP, G/PP/G and PPN cause strong ribosome stalling whereas *e.g.* L/PP/L, CPP or HPP result in a rather weak translational pause. Moreover, the stalling strength is modulated by amino acids located—up to position -5—upstream of the arrest motif [10,12,13]. In this respect, H, K, Q, R or W further pronounce the arrest whereas C, G, L, S or T attenuate it. We therefore define a "polyproline motif" as a consecutive stretch of prolines with flanking residues: $X_{(-2)}X_{(-1)}$-nP-$X_{(+1)}$, n≥2; where $X_{(-2)}$, $X_{(-1)}$ and $X_{(+1)}$ can be any amino acid.

Regardless of the difficulties to translate consecutive proline coding sequences, they occur frequently within prokaryotic and eukaryotic proteomes [14,15]. This in turn implies that the benefits of retaining polyproline motifs significantly outweigh their costs to incorporate them into the nascent polypeptide chain [16]. Proline is unique in terms of being the sole amino acid to adopt *cis* and *trans* conformations, both of which are nearly energetically equal and naturally occur in proteins [17–19]. Notably, a sequence of consecutive prolines results in the formation of either the right-handed poly (*cis*-) proline helix I (PPI) or the left-handed poly (*trans*-) proline helix II (PPII). Beside α-helix and β-sheet, PPII helix is considered to be the third major secondary structure element in proteins and plays an important role in mediating protein-protein and protein-nucleic acid interactions [20–23]. Three consecutive prolines are also an integral part of the active center in the universally conserved Val-tRNA synthetase ValS [14]. The proline triplet in ValS is essential for efficient charging of the tRNA with valine and for preventing mischarging by threonine. These two examples illustrate why nature has evolved a specialized translation elongation factor, referred to as EF-P in bacteria or e/aIF-5A in eukaryotes and archaea, to alleviate ribosome stalling at polyproline motifs [3,5–9,16,24]. The importance of polyproline motifs in proteins is further underlined by the fact that *efp* mutants are characterized by pleiotropic defects. Reportedly, the absence of EF-P impairs bacterial fitness [25,26], membrane integrity [27], motility [28], antibiotic sensitivity [29] and is ultimately lethal for certain bacteria such as *Mycobacterium tuberculosis* [30] and *Neisseria meningitides* [31]. Similarly, IF-5A is an essential protein in archaea [32] as well as in eukaryotes [33] where eIF-5A is associated *e.g.* with cancer [34] and HIV infection [35].

EF-P alleviates polyproline motif-dependent translational arrest, but does not prevent ribosome pausing at these sequences [6,10,36]. The fact that polyproline motifs form a functionally important structural element—the PPII helix—and at the same time interfere with translation poses a major evolutionary conundrum. Are polyproline motifs disfavored during evolution due to their translational burden? Does ribosome stalling caused by polyproline motifs regulate the speed of translation at the protein level in the same way as rare genetic codons and RNA secondary structures cause translational pause at the RNA level [37,38]? To address these questions, we conducted an evolutionary analysis of polyproline motifs in the proteomes of 43

E. coli strains. Our analysis revealed evolutionary selection against polyproline motifs as a consequence of the reduced translation efficiency. Against the overall background of polyproline motif depletion, we observed their frequent occurrence in the vicinity of translational start sites, in the inter-domain regions of multi-domain proteins, and downstream of transmembrane helices, where slow-translating codons are also enriched. This indicates the potential involvement of polyproline motifs in co-translational protein folding and transmembrane helix insertion.

Results

Polyproline motifs are underrepresented in *E. coli* proteomes

We first investigated the overall frequency of polyproline motifs in *E. coli* strains and found 99,386 polyproline motifs within 68,710 (33.3%) proteins from the 43 proteomes considered in this study. Out of these 68,710 proteins, 47,056 proteins (68.5%) harbor only one polyproline motif, 15,027 proteins (21.9%) have two polyproline motifs, and 6,627 proteins (9.6%) have more than 2 polyproline motifs (S1 Fig). We identified 22,253 (22.4%), 21,953 (22.1%) and 55,149 (55.5%) polyproline motifs with strong, medium and weak ribosome stalling effect, respectively. We found that polyproline motifs are significantly underrepresented in all the 43 *E. coli* proteomes compared with randomly generated protein sequences (Fig 1A and S2 Fig; *p*-values $< 2.2e\text{-}16$; average fold change 0.82). Pairs of consecutive prolines show the lowest ratio between the observed and expected frequency (0.84) compared to all other pairs of identical amino acids in *E. coli* K-12 MG1655 (the ratios of the other amino acids are between 1.00 and 1.66, with a mean of 1.17). Moreover, normalized by the random level, the numbers of polyproline motifs negatively depend on the strength of the ribosome pausing effect in all *E. coli* proteomes: in *E. coli* K-12 MG1655, for example, polyproline motifs with strong, medium and weak ribosome stalling effect constitute 55.5%, 70.9% and 104.4% of the random level, respectively (Fig 1B and S3 Fig). Collectively, these findings suggest the existence of evolutionary pressure against ribosome pausing.

To investigate this hypothesis, we grouped the proteins into the core proteome, which encompasses conserved, evolutionary older sequences, and the accessory proteome, which mainly contains proteins of younger origin. Assuming that evolution disfavors polyproline motifs, one would expect them to occur less frequently in the core proteome. Indeed, significantly fewer polyproline motifs were found in proteins belonging to the *E. coli* core set, independent of the arrest strength (Fig 1C and S4 Fig; Mann-Whitney-Wilcoxon test, *p*-values $< 2.2e\text{-}16$; the ratios between motif occurrence of core and accessory proteomes for strong, medium and weak strengths are 0.88, 0.87 and 0.93, respectively). We note that we cannot fully rule out the possibility that this observation is partly due to the differences between the core and accessory proteomes in terms of their functional repertoire (GO terms) and gene expression levels (mean of $\log_{10}$ value: translation efficiency, 1.68 vs 1.57; protein abundance, 2.05 vs 1.86).

We also compared the occurrence of polyproline motifs across the 43 *E. coli* strains. We found that they have the highest occurrence in strain O157:H7 (fold change 0.85) and the lowest occurrence in strain UM146 (fold change 0.79). As seen in S5 Fig, the polyproline motif occurrence in the core proteomes of *E. coli* strains is quite similar (mean fold change 0.76, standard deviation 0.003), while the accessory proteomes are more diverse in this respect (mean fold change 0.85, standard deviation 0.023). We also found that polyproline motif occurrence is highly correlated with the number of proteins in proteomes (S6 Fig; Pearson's r = 0.68, *p*-value = 4.82e-7). By definition, core proteome sizes of *E. coli* strains are the same, while accessory proteome sizes differ. Therefore, this result actually indicates that strains with larger accessory proteomes have more polyproline motifs, as already discussed above.

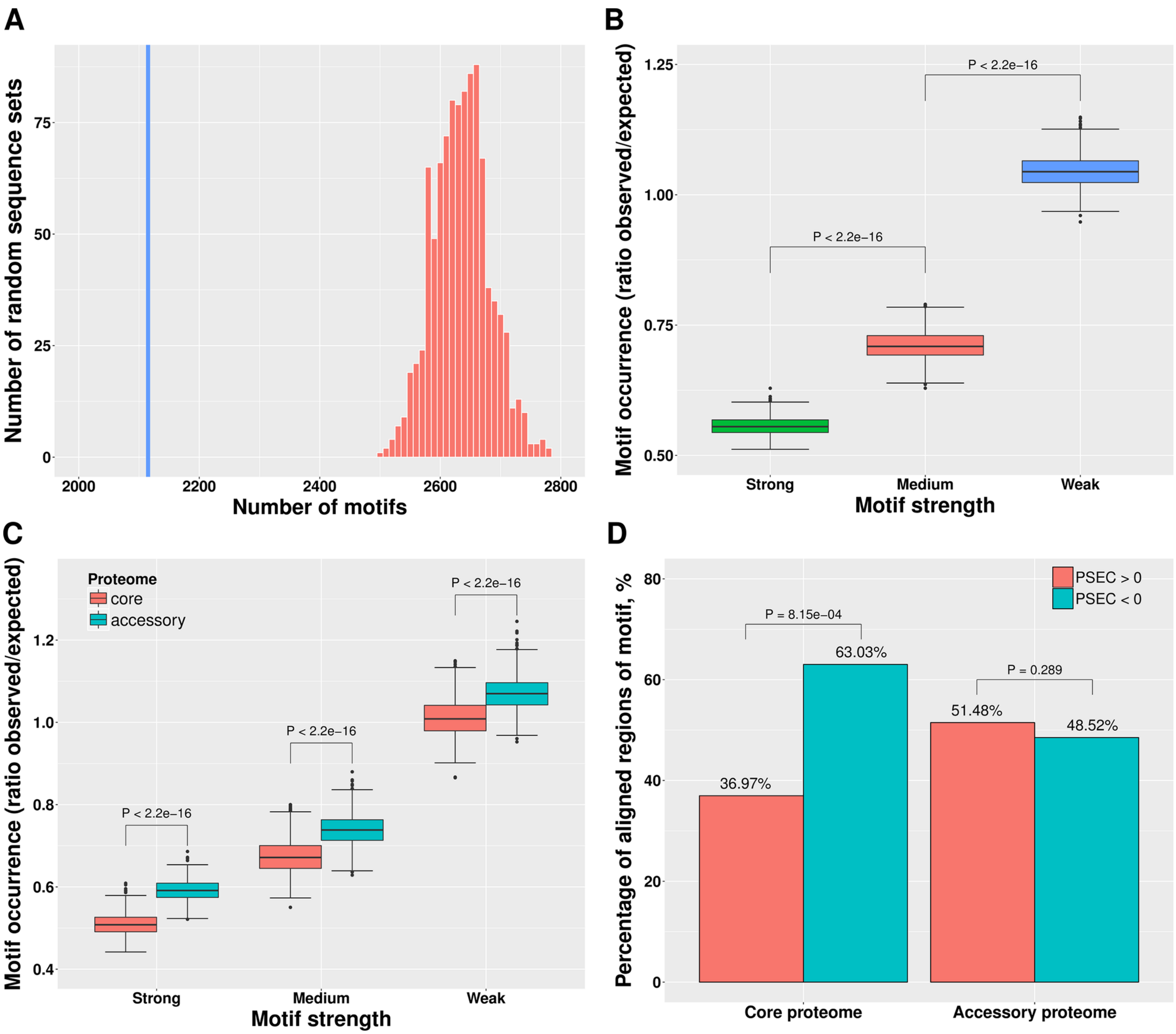

Fig 1. Distribution and conservation of polyproline motifs. (A) Occurrence of polyproline motifs in *E. coli* K-12 MG1655 is lower than the random level (fold change 0.80). The histogram shows the numbers of motifs found in 1,000 sets of random sequences, and the blue line shows the number of motifs found in real sequences. (B) Numbers of polyproline motifs negatively correlate with the strength of the ribosome stalling effect in *E. coli* K-12 MG1655. The differences are significant according to Mann-Whitney-Wilcoxon test. (C) Occurrence of polyproline motifs in the core proteome of *E. coli* K-12 MG1655 is lower than that in the accessory proteome. The differences are significant according to Mann-Whitney-Wilcoxon test. (D) In the core proteome more aligned regions have a negative PSEC (chi-squared test) while in the accessory proteome PSEC values display no strong preference.

https://doi.org/10.1371/journal.pcbi.1005987.g001

Variation of ribosome stalling strength in *E. coli* evolution

We next investigated changes in ribosome stalling strength caused by polyproline motifs in the *E. coli* proteins by considering 3,280 orthologous groups with at least 3 proteins and at least one polyproline motif. Within these orthologous groups, we identified 4,980 aligned regions containing polyproline motifs, of which 1,568 showed changes of the ribosome stalling effect states. Out of the 1,923 evolutionary events 955 were gain events (change from a weaker or no stalling effect state to a stronger state) and 968 were loss events (change from a stronger stalling effect state to a weaker or no stalling effect state). The propensity for stalling effect change (PSEC) was calculated for each of these aligned regions as described in the *Materials and Methods* section. In the core proteome, substantially more aligned regions displayed a negative PSEC (Fig 1D; 63.03% vs 36.97%; chi-squared test, *p*-value = 8.15e-4), indicating that the ribosome stalling effect tends to be lost in evolution. In line with this finding, in the phylogenetically younger accessory proteome, PSEC still displayed no strong preference with 51.48% and 48.52% aligned regions possessing positive and negative PSEC, respectively (Fig 1D; chi-squared test, *p*-value = 0.289). These results are also in line with the notion that evolution generally disfavors polyproline motifs in *E. coli*.

Translational efficiency is the evolutionary driving force for selecting against polyproline motifs

The efficiency of translation and consequently biosynthesis correlates with both translation initiation and elongation rates [39]. Translation elongation rate in turn depends on multiple factors, such as codon bias [36], tRNA levels [40] and the amino acid to be incorporated [2,41], but can also be influenced by an amino acid sequence such as consecutive prolines [5,7,10]. Accordingly, we investigated whether there is a connection between the relative frequency of polyproline motifs and translational efficiency in *E. coli* K-12 MG1655, and found that they are negatively correlated (Fig 2A; Spearman's rho = -0.105, *p*-value = 1.13e-5), which is especially evident in the top 25% of most efficiently translated proteins and for polyproline motifs known to cause a strong translational pause. Occurrence of polyproline motifs also anticorrelates with relative protein abundance (Fig 2B; Spearman's rho = -0.135, *p*-value = 1.47e-8). Thus, in the course of evolution polyproline motifs are more disfavored in those proteins that have a high copy number per cell and need to be efficiently translated, implying a translation efficiency-driven selection pressure against polyproline motifs.

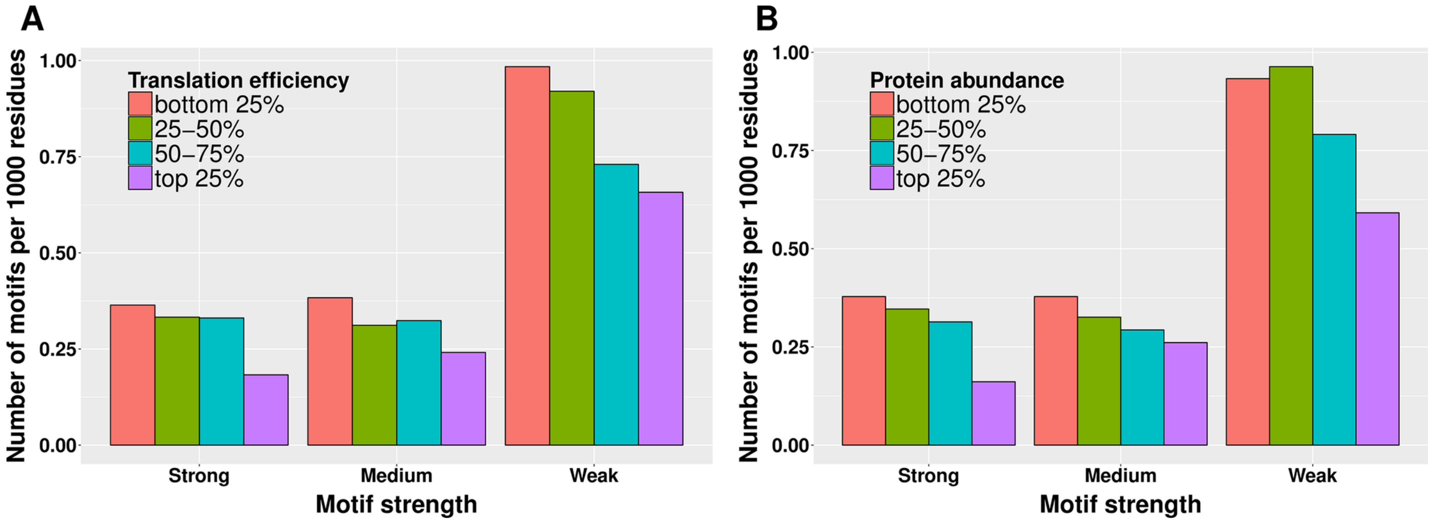

Fig 2. Correlation between translation efficiency, protein abundance and frequency of polyproline motifs. (A) Proteins with high translation efficiency tend to have fewer polyproline motifs (Spearman's rho = -0.105, *p*-value = 1.13e-5). (B) High abundance proteins tend to have fewer motifs (Spearman's rho = -0.135, *p*-value = 1.47e-8).

https://doi.org/10.1371/journal.pcbi.1005987.g002

Polyproline motifs as regulatory elements in protein synthesis

We next investigated whether polyproline-mediated ribosome pausing is exploited in the regulation of translation, focusing on the reference strain *E. coli* K-12 MG1655 comprising 2,115 polyproline motifs in 1,477 proteins (33.9% of the whole proteome) (S1 Table). In 2010, Tuller *et al.* discovered reduced translation efficiency within the first 50 codons of the coding regions [42]. The authors suggested that a slow ramp at the beginning of the ORF might serve as a late stage of translation initiation, being a probate means to reduce ribosomal traffic jams in order to minimize the cost of protein biosynthesis [42–44]. Multiple factors, including slow-translating codons, strong mRNA structures and positively charged amino acids, were implicated in the formation of the ramp [43,44]. We were therefore curious whether there exists an enrichment of polyproline motifs around the start sites of *E. coli* K-12 MG1655 proteins. In the 2,115 polyproline motifs of *E. coli* K-12 MG1655, 325 were found in the first 50 amino acids, and 1,771 located elsewhere in the protein sequence. After normalization by random level, we found a clear enrichment of polyproline motifs in the N-terminal 50 residues (Fig 3A; Mann-Whitney-Wilcoxon test, p-value < 2.2e-16; fold change 0.94 vs 0.78). Thus, similar to the specific codon bias in this region, an accumulation of polyproline motifs might allow adjustment of translational speed in order to minimize the cost of protein production.

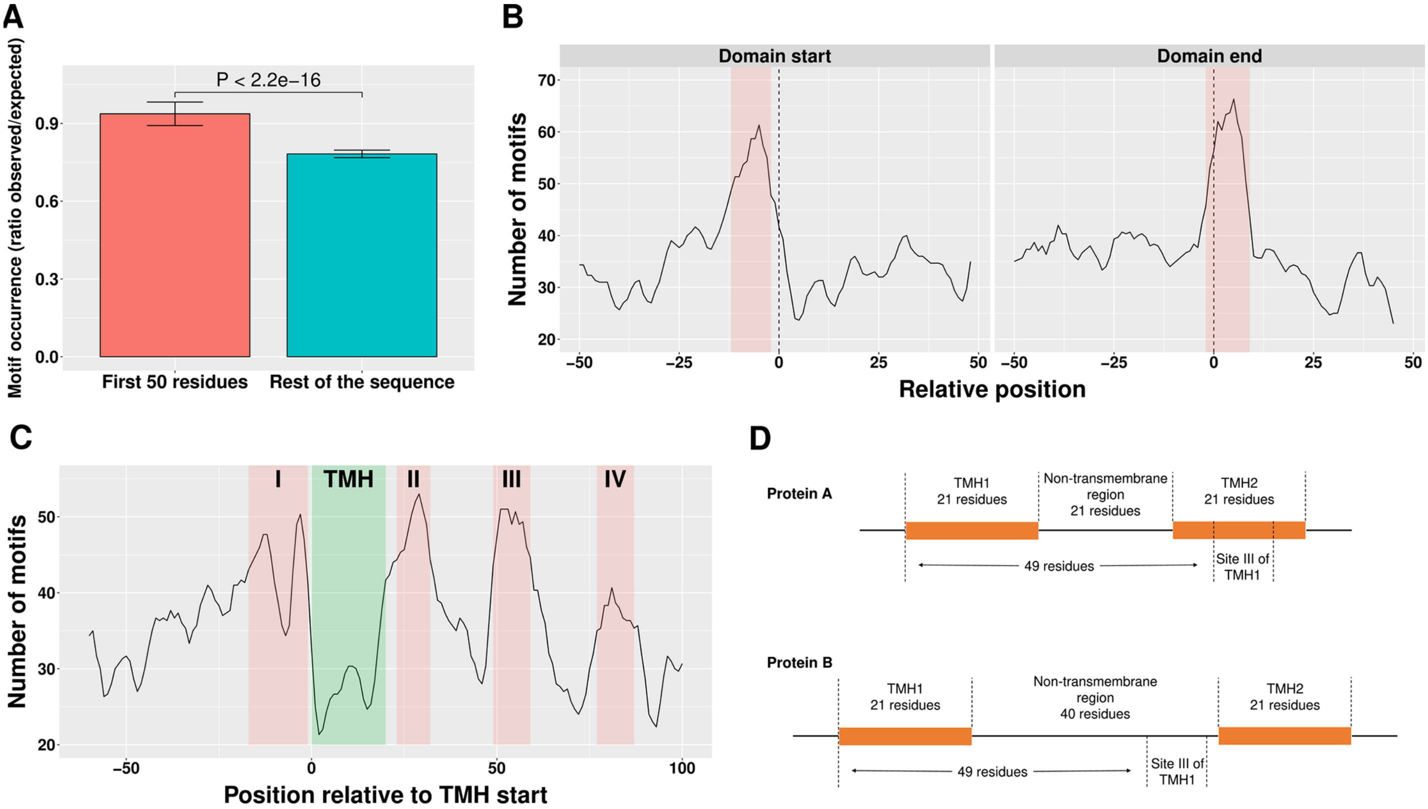

Fig 3. Functional role of polyproline motifs. (A) Occurrence of polyproline motifs in the first 50 residues is higher than elsewhere in the protein sequence (Mann-Whitney-Wilcoxon test, p-value < 2.2e-16; fold change 0.94 vs 0.78). Error bars indicate the standard deviation. (B) Occurrence of polyproline motifs is associated with domain boundaries. Regions with relatively high motif occurrence are marked red. Data are smoothed over a three-residue window. Left: frequency of motifs relative to domain start (dashed line). Right: frequency of motifs relative to domain end (dashed line). The enrichment of motifs in these two regions is significant (p-values < 0.05; fold changes 1.19 and 1.23). (C) Frequency of polyproline motifs relative to the start position of TMH. TMH is marked green (assuming the typical length of 21 residues). Regions with high motif frequency are marked red. Data are smoothed over a three-residue window. (D) Schematic illustration of the site III location relative to TMH and the non-transmembrane region. In protein A site III of TMH1 locates in the TMH2 while in protein B site III of TMH1 is in the non-transmembrane region.

https://doi.org/10.1371/journal.pcbi.1005987.g003

Polyproline motifs coordinate co-translational folding of proteins

Protein folding is a co-translational process, and it is generally believed that structural elements of a protein may influence each other during the folding process [45]. Due to the cooperativity between different parts of the structure, the timing of translation is crucial for proper folding [38]. The non-uniform distribution of synonymous codons with different translation rates fine-tunes the co-translational folding of proteins [46–54]. Fast translation of the mRNA stretches coding for structural domains helps to avoid misfolded intermediates [55], while translational pauses induced by clusters of slow-translating codons in the inter-domain linkers of multi-domain proteins facilitate independent folding of domains to minimize the chance of misfolding [46,53,56–58]. By analogy, we hypothesized that polyproline motifs may coordinate co-translational folding by slowing down translation of inter-domain linkers, and as a consequence, would be expected to occur more frequently between rather than within structural domains.

We therefore investigated the positional preference of polyproline motifs in globular multi-domain proteins. Sequence positions of 7,398 structural domains within 4,080 *E. coli* K-12 MG1655 proteins were obtained from Gene3D database [59]. Out of these proteins, 1,868 (45.8%) are multi-domain proteins possessing the total of 5,186 domains. An inter-domain linker was defined as the sequence span between the boundaries of two consecutive domains (if such a span was shorter than 5 amino acids, it was expanded downstream to achieve the length of 5 amino acids). This procedure yielded 3,318 inter-domain linkers between 5,186 domains.

Indeed, we found that polyproline motifs are significantly depleted in structural domains (p-value = 7.86e-80; fold change 0.56), but not in inter-domain linkers (p-value = 0.912; fold change 1.10). We then investigated the relative location of polyproline motifs with respect to domain boundaries. As seen in Fig 3B, polyproline motifs frequently occur in two regions: 1) -12 to -2 residues relative to the domain start; and 2) -2 to +9 residues relative to the domain end. Polyproline motifs are significantly enriched in these two regions (p-values < 0.05; fold changes for these two regions are 1.19 and 1.23, respectively). Thus, there is a strong correlation between the location of polyproline motifs and the structural domain boundaries, which was also observed for clusters of slow-translating codons [57,58,60,61]. These findings imply that the ribosome stalling effect caused by the polyproline motifs within structural domains may interfere with their folding, while stalling at domain boundaries may facilitate it.

Polyproline motifs facilitate co-translational insertion of transmembrane helices

Another typical co-translational process is the targeting of α-helical transmembrane proteins (TPs) to the translocons, mediated by the signal recognition particle (SRP), and their insertion into the membrane [62,63]. This process has been found to be facilitated by translational pause [50,64–69]. A recent study by Fluman *et al.* identified two translation pauses, triggered by Shine-Dalgarno-like elements in *E. coli* mRNAs, that contribute to the SRP-mediated targeting of TPs [64]. The first pause occurs before the nascent peptide emerges from the exit tunnel of the ribosome (16 to 30 codons of the protein) and the second one occurs after the emergence of the first transmembrane helix (TMH) (-5 to +1 codons relative to the start of the second TMH). In the fungus *Emericella nidulans*, Dessen *et al.* identified two translational pauses occurring at the distance of approximately 45 and 70 codons from TMHs, caused by clusters of slow-translating codons and presumed to facilitate translocon-mediated co-translational insertion of TMH [66].

We investigated the occurrence and location of polyproline motifs in TPs. Based on the UniProt [70] annotation, we identified 912 TPs from *E. coli* K-12 containing the total of 5,672 TMHs. We found that 39.3% (358) of these TPs harbor polyproline motifs, which is even higher than the percentage of soluble proteins (32.6%; chi-squared test, *p*-value = 1.6e-4). No enrichment of polyproline motifs around the pause sites identified by Fluman *et al.* was observed. However, as seen in Fig 3C, we found that i) polyproline motifs rarely occur within TMH; and ii) polyproline motifs display a relatively high occurrence in four positions (positions -17 to -1, 23 to 32, 49 to 59 and 77 to 87 relative to TMH start; termed here site I, II, III and IV, respectively). The depletion of polyproline motifs in TMH is significant (*p*-value = 1.65e-27; fold change 0.39) implying that the ribosome stalling effect caused by the polyproline motifs may interfere with the folding of TPs. It should be noted that the site positions are shown relative to the start of a TMH, and thus in some cases the given region can actually be located in another TMH (see Fig 3D for illustration). We therefore tested the enrichment/depletion of the polyproline motifs in each of the four sites described above separately in TMH and in non-transmembrane regions. For example, out of the 4,439 site IV regions 3,013 and 1,426 regions are located in TMH and non-transmembrane regions, respectively. For all four sites, significant depletion of polyproline motifs was evident in TMH regions (*p*-values for sites I, II, III and IV are 2.63e-6, 1.86e-3, 7.84e-3 and 1.98e-3, respectively; fold changes of these 4 sites are 0.57, 0.56, 0.71 and 0.64, respectively), while in non-transmembrane regions a significant enrichment of polyproline motifs was observed for site III (*p*-value = 0.035; fold change 1.39). The location of this site is similar to the location of one of the translational pauses (approximately 45 codons from TMHs) identified by Dessen *et al.* Considering that most of the TMHs are 21 residues in length (S7 Fig) and that about 28 amino acids can be accommodated in the ribosome exit tunnel [71], ribosome stalling at site III may occur after the TMH has emerged from the ribosome exit tunnel and is being inserted into the membrane by translocon [63,72]. We therefore speculate that the translational pause at site III could provide a time delay for the efficient insertion of TMH.

Discussion

Proline is a poor substrate for the ribosomal peptidyl transfer reaction [2–4], and consecutive prolines cause ribosome stalling [5]. The bacterial elongation factor P (EF-P) and its eukaryotic and archaeal orthologs e/aIF5A alleviate this stalling to some degree, but cannot fully compensate the translational burden imposed by polyproline motifs [6–8,10]. The presence of a large number of such motifs in bacterial proteomes might imply their biological significance, yet their precise functional role remains poorly understood [14,15].

In this study, we made a comprehensive attempt to shed light on the functional role of polyproline motifs by investigating their distribution and evolution in the proteomes of 43 *E. coli* strains. We found evidence of evolutionary selection pressure against translational stalling caused by polyproline motifs. Translational efficiency and protein abundance negatively correlate with the frequency of polyproline motifs and thus might be the driving force for their loss. Against the general trend of losing polyproline motifs during the course of evolution, we observed accumulation of polyproline motifs close to the protein N-terminus, in inter-domain regions of multi-domain proteins as well as downstream of transmembrane helices. We therefore speculate that the time gain caused by translational pause at polyproline motifs might be crucial for translational regulation, domain folding, and the proper membrane insertion, respectively.

Evolutionary selection for high efficiency of protein synthesis is one of the forces shaping mRNA sequences. For example, unequal usage of synonymous codons reflects an adaption of the codon usage to the available tRNA pool, with slow-translating codons used much more rarely than fast-translating codons [73,74]. However, protein sequence elements were also found to influence the translation rate by interacting with the ribosome exit tunnel or impairing the peptidyl transfer reaction [7,75]. An important question, which arises in this context, is whether there exists protein-level evolutionary selection for high translation efficiency. Recently, Tuller *et al*. found that short peptides, which induce ribosome stalling in yeast by interacting with the ribosomal exit tunnel, tend to be either over or underrepresented in the proteome [76]. They hypothesized that short peptide sequences were under evolutionary selection based on their synthetic efficiency. Our results show that polyproline motifs, which induce ribosome stalling by slowing down the peptidyl transfer reaction, are significantly underrepresented in *E. coli* proteomes, and that selection is more evident against motifs causing stronger ribosome stalling and in proteins with higher translation efficiency. These findings support the conjecture that translation efficiency-based evolutionary pressure shapes protein sequences.

Against the overall background of polyproline motif depletion, our investigation of the intra-molecular distribution pattern of polyproline motifs revealed their overrepresentation at several strategic locations, indicating their regulatory role in translation elongation. Translation elongation is a non-uniform process, which is subject to strict regulation [1,16] both in terms of the quantity of the translation products [77] and the intra-molecular variation of the elongation rate, which ensures the quality of the synthesized proteins by coordinating co-translational processes [47,64,78]. The role of polyproline motifs in the regulation of the overall translation elongation rate is exemplified by the lysine-dependent acid stress response regulator CadC of *E. coli* [7,16,79]. This membrane-integrated pH-sensor and transcriptional activator contains two polyproline motifs, which allow for fine-tuning of its copy number. The amount of the CadC protein is crucial for regulating the expression of the target operon. Analogously, precisely regulated translational output of the polyproline-containing receptor CpxA is required for *Shigella flexneri* virulence [80]. The intra-molecular variation of the elongation rate has so far been thought to be regulated by *cis*-acting elements embedded in the translated mRNA [57,81], such as clusters of slow-translating codons [38] and Shine-Dalgarno-like RNA sequences [64] (although the latter notion has recently been challenged [82]), as well as by *trans*-acting molecules, such as the signal recognition particle, which arrests translation elongation while targeting proteins to the membrane [83,84]. Our study highlights the role of polyproline motifs in coordinating the co-translational protein folding and transmembrane helix insertion, implying that they could serve as protein-level *cis*-acting elements, which directly regulate the rate of translation elongation.

The phenomenon we observed is not specific to *E. coli*. We also investigated the occurrence of polyproline motifs in *Bacillus subtilis*, a Gram-positive bacterium, and obtained qualitatively similar results (S8 Fig and S9 Fig). We therefore speculate that the polyproline motif-mediated regulation of translation elongation may be universal in bacteria. *B. subtilis* has a lower occurrence of polyproline motifs than *E. coli* (average fold change 0.73 vs 0.82), which may indicate a stronger evolutionary selection against polyproline motifs. We initiated a follow-up study of polyproline motif occurrence in bacteria and also in eukaryotes, although the ribosome arresting sequences in eukaryotes are not limited to consecutive prolines [85,86]. In another follow up study we are investigating the interplay between the RNA level elements, such as codon usage and RNA structure, and polyproline motifs (Qi, F. *et al*., in preparation).

Materials and methods

Proteomes and orthologous groups of *E. coli*

We obtained *Escherichia coli* proteomes and orthology assignments from the OMA database [87]. The total of 206,360 protein sequences from six out of seven *E. coli* phylotypes [88] were downloaded (S2 Table). We also obtained 11,356 orthologous groups covering 195,056 proteins.

The core- and accessory proteomes of *E. coli*

The core- and accessory proteomes were defined based on the occurrence of orthologous groups. An orthologous group was classified as belonging to the core proteome if it was present in all the 43 *E. coli* proteomes, otherwise it was considered to belong to the accessory proteome. All proteins not assigned to any orthologous group were classified as belonging to the accessory proteome. This procedure yielded a core proteome of *E. coli* covering 73,745 proteins and an accessory proteome covering 132,615 proteins.

Identification of polyproline motifs in real and random sequences

Using the program *fuzzpro* from the EMBOSS package [89] we identified polyproline motifs in the *E. coli* proteins. The same procedure was applied to randomly generated sequences. Each amino acid sequence in our dataset was shuffled 1,000 times while maintaining its composition using the program *shuffleseq* from the EMBOSS package [89], yielding 1,000 sets of random *E. coli* protein sequences.

Enrichment and depletion of polyproline motifs

We used the SPatt algorithm [90] to assess the enrichment and depletion of polyproline motifs, taking into account occurrence patterns of proline in various parts of protein structure. SPatt determines the expected occurrence of a sequence motif based on a Markov chain model of order m (model Mm), compares the observed occurrence with expected one, and calculates the p-value for the significance of a motif's enrichment or depletion. Choosing a model Mm means taking into account the m-mer and (m+1)-mer compositions while determining the expected occurrence. For example, the model M0 solely takes into account the amino acid composition, while choosing the model M1 takes into account the compositions of amino acid monomers and dimers. For a motif of length l, the maximum m is (l-2). In our case, although a polyproline motif can have more than 2 residues, the essential part of a polyproline motif is the proline stretch with at least two consecutive proline residues. Therefore, we chose model M0 in our tests.

Normalization of polyproline motif occurrence

The occurrence of polyproline motifs in proteins was normalized by the polyproline motif occurrence in randomly generated sequences. Each amino acid sequence (either full protein sequences or specific sequence segments of interest) was shuffled 1,000 times while maintaining its composition using the program *shuffleseq* from the EMBOSS package [89], yielding 1,000 sets of random sequences. The number of times the polyproline motif occurred in a real sequence was then divided by the number of times the same motif occurred in each of the 1,000 random sequences, yielding a vector of 1,000 ratios between the observed and the expected polyproline motif occurrence. The Mann-Whitney-Wilcoxon test was employed to assess the significance of the difference between two such vectors corresponding to two different sequences or sequence segments. This procedure was carried out for each strain of *E. coli* separately.

Fold change of polyproline motif occurrence

The fold change of polyproline motif occurrence is used as a measure of the enrichment/depletion level of polyproline motifs. It is defined as the ratio between the observed and expected occurrence of polyproline motifs, and is calculated as:

$$\text{Fold\_change} = \frac{N_{obs}}{N_{exp}} \tag{1}$$

Where N_{obs} and N_{exp} are the observed and expected occurrences of polyproline motifs, respectively. The N_{exp} is either the mean value of the polyproline motif occurrences in 1,000 sets of random sequences (for fold change of whole proteomes; see *Normalization of polyproline motif occurrence* for detail) or the mean of the distribution of expected motif occurrence calculated by SPatt algorithm (for fold change of structural domains, domain linkers, TMHs and regions bracketing domains and TMHs; see *Enrichment and depletion of polyproline motifs* for detail).

Classification of polyproline motifs

Polyproline motifs were classified into three groups (strong, medium and week) according to their predicted ribosomal translation arrest strength. The prediction is based on experimental data both from systematic *in vitro* and *in vivo* analyses [10,12,14–16] (S3 Table and S4 Table).

As described in the *Introduction* section, the ribosome stalling strength of a $X_{(-2)}X_{(-1)}$-nP-$X_{(+1)}$ motif is dependent on the number of consecutive prolines and on the flanking amino acids. First, we classified the flanking residues $X_{(-2)}$, $X_{(-1)}$ and $X_{(+1)}$ (motifs involving ambiguous amino acids were excluded from consideration) according to their influence on the ribosome stalling strength. If a flanking residue of the polyproline motif in an *E. coli* strain lacking *efp* (Δ*efp*) was responsible for a decrease of the translational output by ≥70% compared to a wildtype control, the residue was defined as strong [5,10,12]. In cases where the protein synthesis was reduced by 30–60%, we classified the stalling strength as medium. In all other cases, the polyproline sequence context was assumed to cause only a weak arrest. All possible $X_{(-2)}X_{(-1)}$-nP-$X_{(+1)}$ motifs and their respective arrest strength are listed in S4 Table and S5 Table. Based on our classification, we correlated the predicted motif strength to available ribosome profiling data [10]. Woolstenhulme *et al.* compared the ribosome occupancy at a diprolyl motif with the occupancy downstream of the motif in an Δ*efp* strain [10]. Stalling was ranked according to the observed assymetry (ratio) between these two values. When setting an assymetry quotient of 2.00 as a threshold for proteins subject to strong translation arrest, we found that more than 75% of them possess at least one medium or strong polyproline motif. This number further increases to ~80% and ~90% when applying more stringent cutoffs of 3.00 and 5.00 to the assymetry score, respectively (S6 Table).

Word frequency in protein sequences

Frequencies of each single and dimer amino acid in protein sequences were calculated using the *compseq* program from the EMBOSS package [89]. For each amino acid dimer, an expected frequency was additionally calculated based on the observed frequencies of single amino acids.

Multiple alignment of protein sequences

Multiple alignments of protein sequences in each orthologous group were computed using the Clustal Omega software [91] with all default parameters.

Construction of phylogenetic trees

Phylogenetic trees for each orthologous group with at least three proteins containing at least one polyproline motif were reconstructed using the PhyML software [92]. These trees were then rooted at midpoint.

Reconstruction of evolutionary events

In order to reconstruct the gain and loss of the ribosome stalling effect in the evolutionary history of *E. coli* protein families, we first assigned one of the four possible ribosome stalling states [S (strong), M (medium), W (weak) and N (none)] to all the exterior nodes (leaves) of the phylogenetic trees. Subsequently, the Maximum Likelihood algorithm [93] was employed to reconstruct the states of ancestral nodes (internal nodes). The change of state between a given node and its ancestral node from a stronger stalling effect state to a weaker or no stalling effect state was defined as a loss of the stalling effect, while the change from a weaker or no stalling effect state to a stronger state was defined as a gain event.

Propensity of stalling effect change

We defined propensity of stalling effect change (PSEC) similar to propensity of gene loss (PGL) frequently used in evolutionary studies [94]. PGL captures the idea that the longer the time during which a gene could have been lost but was not, the lower the propensity of this gene to be lost. PGL is thus defined as the ratio between the total length of branches in which the gene is lost and the total length of branches in which the gene could have been lost [95,96]. Similarly, PSEC captures the idea that the longer the time the stalling effect of a motif could have been gained/lost but was not, the lower the propensity of the stalling effect to be gained/lost. However, our model is somewhat more complex than the PGL model, since the PGL only considers gene loss and we have to consider both gain and loss of the stalling effect. Therefore, the PSEC is calculated as the difference between the propensities of gain and loss of the stalling effect:

$$\mathrm{PSEC} = \frac{\sum B_g}{\sum B_{cg}} - \frac{\sum B_l}{\sum B_{cl}} \tag{2}$$

where B_g and B_l are the lengths of the branches in which the stalling effect was gained and lost, respectively, and B_{cg} and B_{cl} are the lengths of branches in which the stalling effect could have been gained and lost, respectively. Thus, a positive PSEC indicates that the stalling effect of a sequence motif tends to be gained, while a negative PSEC indicates that it tends to be lost during evolution.

Protein abundance, gene expression, and translation efficiency

Protein abundance data used in this study was from [97,98], covering 2,163 proteins. Microarray data on transcription levels of 2,710 genes from *E. coli* K-12 MG1655 under standard growth conditions was downloaded from the ASAP database [99]. Translation efficiency for each of the 1,743 genes present in both datasets was calculated as:

$$Translation\_efficiency_i = \frac{Protein\_abundance_i}{Transcription\_level_i} \tag{3}$$

Domain composition of the *E. coli* proteins

Sequence positions of 7,398 structural domains in 4,080 *E. coli* K-12 MG1655 proteins were obtained from the Gene3D database [59].

Transmembrane segments

We obtained the sequence positions of 5,672 transmembrane segments within 912 α-helical transmembrane proteins from the UniProt database [70]. Since reviewed data on transmembrane proteins of *E. coli* K-12 MG1655 (taxonomy ID 511145) are not available in the UniProt database, we used the reviewed data of *E. coli* K-12 (taxonomy ID 83333) instead.

Supporting information

S1 Fig. Numbers of *E. coli* proteins with 1, 2 and >2 polyproline motifs.
(PDF)

S2 Fig. Occurrence of polyproline motifs is lower than the random level. The histogram shows the numbers of motifs found in 1,000 sets of random sequences and the blue line shows the number of motifs found in real sequences. The results for 42 *E. coli* strains (except for *E. coli* K-12 MG1655) are shown. The OMA id and fold change for each strain are shown in the panel title. For mapping OMA ids to names of strains, please see S2 Table.
(PDF)

S3 Fig. Numbers of polyproline motifs negatively correlate with the strength of the ribosome stalling effect. The results for 42 *E. coli* strains (except for *E. coli* K-12 MG1655) are shown. The OMA id of each strain is shown in the panel title. For mapping OMA ids to names of strains, please see S2 Table. All the differences are significant according to Mann-Whitney-Wilcoxon test, *p*-values < 2.2e-16.
(PDF)

S4 Fig. Occurrence of polyproline motifs in the core proteome is lower than that in the accessory proteome. The results for 42 *E. coli* strains (except for *E. coli* K-12 MG1655) are shown. The OMA id of each strain is shown in the panel title. For mapping OMA ids to names of strains, please see S2 Table. All the differences are significant according to Mann-Whitney-Wilcoxon test, *p*-values < 2.2e-16.
(PDF)

S5 Fig. Occurrences of polyproline motifs in the core/accessory proteome in the 43 *E. coli* strains. The OMA id of each strain is shown on the x-axis. For mapping OMA ids to the names of strains, please see S2 Table. Fold changes for the core proteome: mean 0.76, standard deviation 0.003. Fold changes for the accessory proteome: mean 0.85. standard deviation 0.023.
(PDF)

S6 Fig. Occurrence of polyproline motifs is correlated with the number of proteins in the proteomes. Pearson's r = 0.68, *p*-value = 4.82e-7.
(PDF)

S7 Fig. Histogram of the TMH length.
(PDF)

S8 Fig. Occurrence of polyproline motifs in *B. subtilis* is lower than the random level. The histogram shows the numbers of motifs found in 1,000 sets of random sequences and the blue line shows the number of motifs found in real sequences. The name and fold change of each strain are shown in the panel title. The proteomes of the 4 *B. subtilis* strains including 16,678 proteins were downloaded from the OMA database [87].
(PDF)

S9 Fig. The distribution of polyproline motifs in *B. subtilis* is qualitatively similar to *E. coli*. The strain 168 were used as a reference strain of *B. subtilis*. (A) Occurrence of polyproline motifs in the first 50 residues is higher than elsewhere in the protein sequence (Mann-Whitney-Wilcoxon test, *p*-value < 2.2e-16; fold change 0.82 vs 0.73). Error bars indicate the

standard deviation. (B) Occurrence of polyproline motifs is associated with domain boundaries. Regions with relatively high motif occurrence are marked red. Data are smoothed over a three-residue window. Left: frequency of motifs relative to domain start (dashed line). Right: frequency of motifs relative to domain end (dashed line). Sequence positions of 6,086 structural domains in 3,076 *B. subtilis* strain 168 proteins were obtained from the Gene3D database [59]. (C) Frequency of polyproline motifs relative to the start position of TMH. TMH is marked green (assuming the typical length of 21 residues). Regions with high motif frequency are marked red. Data are smoothed over a three-residue window. The sequence positions of 5,456 transmembrane segments within 984 α-helical transmembrane proteins of *B. subtilis* strain 168 were obtained from the UniProt database [70]. (D) The typical length of TMHs in *B. subtilis* proteins is 21 residues.
(PDF)

S1 Table. The list of 2,115 polyproline motifs identified in *E. coli* K-12 MG1655.
(XLSX)

S2 Table. List of all 43 *E. coli* strains used in this study.
(XLSX)

S3 Table. Classification of the effect amino acids at position $X_{(-2)}$, $X_{(-1)}$ and $X_{(+1)}$ exert on ribosomal stalling strength.
(PDF)

S4 Table. Rules for the prediction of ribosomal stalling strength induced by a $X_{(-2)}X_{(-1)}$-nP-$X_{(+1)}$ motif.
(PDF)

S5 Table. Motif classification.
(XLSX)

S6 Table. Matching the predicted stalling strength of polyproline motifs with the ribosome profiling data from C. J. Woolstenhulme [Cell Rep 11:13–21, 2015].
(XLSX)

Acknowledgments

We thank Peter Hönigschmid and Jianfeng Sun for constructive discussions.

Author Contributions

Conceptualization: Kirsten Jung, Jürgen Lassak, Dmitrij Frishman.

Formal analysis: Fei Qi.

Investigation: Fei Qi.

Methodology: Fei Qi, Magdalena Motz, Kirsten Jung, Jürgen Lassak, Dmitrij Frishman.

Project administration: Kirsten Jung, Jürgen Lassak, Dmitrij Frishman.

Supervision: Kirsten Jung, Jürgen Lassak, Dmitrij Frishman.

Visualization: Fei Qi.

Writing – original draft: Fei Qi, Magdalena Motz, Kirsten Jung, Jürgen Lassak, Dmitrij Frishman.

Writing – review & editing: Fei Qi, Magdalena Motz, Kirsten Jung, Jürgen Lassak, Dmitrij Frishman.

References

1. Varenne S, Buc J, Lloubes R, Lazdunski C. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J Mol Biol. 1984; 180: 549–76. PMID: 6084718
2. Pavlov MY, Watts RE, Tan Z, Cornish VW, Ehrenberg M, Forster AC. Slow peptide bond formation by proline and other N-alkylamino acids in translation. Proc Natl Acad Sci U S A. 2009; 106: 50–4. https://doi.org/10.1073/pnas.0809211106 PMID: 19104062
3. Doerfel LK, Wohlgemuth I, Kothe C, Peske F, Urlaub H, Rodnina M V. EF-P is essential for rapid synthesis of proteins containing consecutive proline residues. Science. 2013; 339: 85–8. https://doi.org/10.1126/science.1229017 PMID: 23239624
4. Doerfel LK, Wohlgemuth I, Kubyshkin V, Starosta AL, Wilson DN, Budisa N, et al. Entropic Contribution of Elongation Factor P to Proline Positioning at the Catalytic Center of the Ribosome. J Am Chem Soc. 2015; 137: 12997–3006. https://doi.org/10.1021/jacs.5b07427 PMID: 26384033
5. Peil L, Starosta AL, Lassak J, Atkinson GC, Virumäe K, Spitzer M, et al. Distinct XPPX sequence motifs induce ribosome stalling, which is rescued by the translation elongation factor EF-P. Proc Natl Acad Sci U S A. 2013; 110: 15265–70. https://doi.org/10.1073/pnas.1310642110 PMID: 24003132
6. Tanner DR, Cariello DA, Woolstenhulme CJ, Broadbent MA, Buskirk AR. Genetic identification of nascent peptides that induce ribosome stalling. J Biol Chem. 2009; 284: 34809–18. https://doi.org/10.1074/jbc.M109.039040 PMID: 19840930
7. Ude S, Lassak J, Starosta AL, Kraxenberger T, Wilson DN, Jung K. Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches. Science. 2013; 339: 82–5. https://doi.org/10.1126/science.1228985 PMID: 23239623
8. Gutierrez E, Shin B-S, Woolstenhulme CJ, Kim J- R, Saini P, Buskirk AR, et al. eIF5A promotes translation of polyproline motifs. Mol Cell. 2013; 51: 35–45. https://doi.org/10.1016/j.molcel.2013.04.021 PMID: 23727016
9. Hersch SJ, Wang M, Zou SB, Moon K-M, Foster LJ, Ibba M, et al. Divergent protein motifs direct elongation factor P-mediated translational regulation in Salmonella enterica and Escherichia coli. MBio. 2013; 4: e00180–13. https://doi.org/10.1128/mBio.00180-13 PMID: 23611909
10. Woolstenhulme CJ, Guydosh NR, Green R, Buskirk AR. High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. Cell Rep. 2015; 11: 13–21. https://doi.org/10.1016/j.celrep.2015.03.014 PMID: 25843707
11. Cymer F, Hedman R, Ismail N, von Heijne G. Exploration of the arrest peptide sequence space reveals arrest-enhanced variants. J Biol Chem. 2015; 290: 10208–15. https://doi.org/10.1074/jbc.M115.641555 PMID: 25713070
12. Starosta AL, Lassak J, Peil L, Atkinson GC, Virumäe K, Tenson T, et al. Translational stalling at polyproline stretches is modulated by the sequence context upstream of the stall site. Nucleic Acids Res. 2014; 42: 10711–9. https://doi.org/10.1093/nar/gku768 PMID: 25143529
13. Elgamal S, Katz A, Hersch SJ, Newsom D, White P, Navarre WW, et al. EF-P dependent pauses integrate proximal and distal signals during translation. PLoS Genet. 2014; 10: e1004553. https://doi.org/10.1371/journal.pgen.1004553 PMID: 25144653
14. Starosta AL, Lassak J, Peil L, Atkinson GC, Woolstenhulme CJ, Virumäe K, et al. A conserved proline triplet in Val-tRNA synthetase and the origin of elongation factor P. Cell Rep. 2014; 9: 476–83. https://doi.org/10.1016/j.celrep.2014.09.008 PMID: 25310979
15. Mandal A, Mandal S, Park MH. Genome-wide analyses and functional classification of proline repeat-rich proteins: potential role of eIF5A in eukaryotic evolution. PLoS One. 2014; 9: e111800. https://doi.org/10.1371/journal.pone.0111800 PMID: 25364902
16. Lassak J, Wilson DN, Jung K. Stall no more at polyproline stretches with the translation elongation factors EF-P and IF-5A. Mol Microbiol. 2016; 99: 219–35. https://doi.org/10.1111/mmi.13233 PMID: 26416626
17. Lu KP, Finn G, Lee TH, Nicholson LK. Prolyl cis-trans isomerization as a molecular timer. Nat Chem Biol. 2007; 3: 619–29. https://doi.org/10.1038/nchembio.2007.35 PMID: 17876319
18. Thapar R. Roles of Prolyl Isomerases in RNA-Mediated Gene Expression. Biomolecules. 2015; 5: 974–99. https://doi.org/10.3390/biom5020974 PMID: 25992900
19. Macinga DR, Parojcic MM, Rather PN. Identification and analysis of aarP, a transcriptional activator of the 2'-N-acetyltransferase in Providencia stuartii. J Bacteriol. 1995; 177: 3407–13. PMID: 7768849

20. Adzhubei AA, Sternberg MJE, Makarov AA. Polyproline-II helix in proteins: structure and function. J Mol Biol. 2013; 425: 2100–32. https://doi.org/10.1016/j.jmb.2013.03.018 PMID: 23507311

21. Adzhubei AA, Eisenmenger F, Tumanyan VG, Zinke M, Brodzinski S, Esipova NG. Third type of secondary structure: noncooperative mobile conformation. Protein Data Bank analysis. Biochem Biophys Res Commun. 1987; 146: 934–8. Available: http://www.ncbi.nlm.nih.gov/pubmed/3619942 PMID: 3619942

22. Adzhubei AA, Eisenmenger F, Tumanyan VG, Zinke M, Brodzinski S, Esipova NG. Approaching a complete classification of protein secondary structure. J Biomol Struct Dyn. 1987; 5: 689–704. https://doi.org/10.1080/07391102.1987.10506420 PMID: 3271488

23. Jha AK, Colubri A, Zaman MH, Koide S, Sosnick TR, Freed KF. Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library. Biochemistry. 2005; 44: 9691–702. https://doi.org/10.1021/bi0474822 PMID: 16008354

24. Doerfel LK, Rodnina M V. Elongation factor P: Function and effects on bacterial fitness. Biopolymers. 2013; 99: 837–45. https://doi.org/10.1002/bip.22341 PMID: 23828669

25. Yanagisawa T, Sumida T, Ishii R, Takemoto C, Yokoyama S. A paralog of lysyl-tRNA synthetase aminoacylates a conserved lysine residue in translation elongation factor P. Nat Struct Mol Biol. 2010; 17: 1136–43. https://doi.org/10.1038/nsmb.1889 PMID: 20729861

26. Lassak J, Keilhauer EC, Fürst M, Wuichet K, Gödeke J, Starosta AL, et al. Arginine-rhamnosylation as new strategy to activate translation elongation factor P. Nat Chem Biol. 2015; 11: 266–70. https://doi.org/10.1038/nchembio.1751 PMID: 25686373

27. Zou SB, Hersch SJ, Roy H, Wiggers JB, Leung AS, Buranyi S, et al. Loss of elongation factor P disrupts bacterial outer membrane integrity. J Bacteriol. 2012; 194: 413–25. https://doi.org/10.1128/JB.05864-11 PMID: 22081389

28. Kearns DB, Chu F, Rudner R, Losick R. Genes governing swarming in Bacillus subtilis and evidence for a phase variation mechanism controlling surface motility. Mol Microbiol. 2004; 52: 357–69. https://doi.org/10.1111/j.1365-2958.2004.03996.x PMID: 15066026

29. Rajkovic A, Erickson S, Witzky A, Branson OE, Seo J, Gafken PR, et al. Cyclic Rhamnosylated Elongation Factor P Establishes Antibiotic Resistance in Pseudomonas aeruginosa. MBio. 2015; 6: e00823. https://doi.org/10.1128/mBio.00823-15 PMID: 26060278

30. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol. 2003; 48: 77–84. PMID: 12657046

31. Yanagisawa T, Takahashi H, Suzuki T, Masuda A, Dohmae N, Yokoyama S. Neisseria meningitidis Translation Elongation Factor P and Its Active-Site Arginine Residue Are Essential for Cell Viability. PLoS One. 2016; 11: e0147907. https://doi.org/10.1371/journal.pone.0147907 PMID: 26840407

32. Gäbel K, Schmitt J, Schulz S, Näther DJ, Soppa J. A comprehensive analysis of the importance of translation initiation factors for Haloferax volcanii applying deletion and conditional depletion mutants. PLoS One. 2013; 8: e77188. https://doi.org/10.1371/journal.pone.0077188 PMID: 24244275

33. Dever TE, Gutierrez E, Shin B-S. The hypusine-containing translation factor eIF5A. Crit Rev Biochem Mol Biol. 2014; 49: 413–25. https://doi.org/10.3109/10409238.2014.939608 PMID: 25029904

34. Sievert H, Pällmann N, Miller KK, Hermans-Borgmeyer I, Venz S, Sendoel A, et al. A novel mouse model for inhibition of DOHH-mediated hypusine modification reveals a crucial function in embryonic development, proliferation and oncogenic transformation. Dis Model Mech. 2014; 7: 963–76. https://doi.org/10.1242/dmm.014449 PMID: 24832488

35. Hauber I, Bevec D, Heukeshoven J, Krätzer F, Horn F, Choidas A, et al. Identification of cellular deoxyhypusine synthase as a novel target for antiretroviral therapy. J Clin Invest. 2005; 115: 76–85. https://doi.org/10.1172/JCI21949 PMID: 15630446

36. Chevance FF V, Le Guyon S, Hughes KT. The effects of codon context on in vivo translation speed. PLoS Genet. 2014; 10: e1004392. https://doi.org/10.1371/journal.pgen.1004392 PMID: 24901308

37. Chen C, Zhang H, Broitman SL, Reiche M, Farrell I, Cooperman BS, et al. Dynamics of translation by single ribosomes through mRNA secondary structures. Nat Struct Mol Biol. Nature Publishing Group; 2013; 20: 582–588. https://doi.org/10.1038/nsmb.2544 PMID: 23542154

38. Marin M. Folding at the rhythm of the rare codon beat. Biotechnol J. 2008; 3: 1047–57. https://doi.org/10.1002/biot.200800089 PMID: 18624343

39. Quax TEF, Claassens NJ, Söll D, van der Oost J. Codon Bias as a Means to Fine-Tune Gene Expression. Mol Cell. 2015; 59: 149–61. https://doi.org/10.1016/j.molcel.2015.05.035 PMID: 26186290

40. Dana A, Tuller T. The effect of tRNA levels on decoding times of mRNA codons. Nucleic Acids Res. 2014; 42: 9171–81. https://doi.org/10.1093/nar/gku646 PMID: 25056313

41. Charneski CA, Hurst LD. Positively charged residues are the major determinants of ribosomal velocity. PLoS Biol. 2013; 11: e1001508. https://doi.org/10.1371/journal.pbio.1001508 PMID: 23554576

42. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, et al. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell. 2010; 141: 344–54. https://doi.org/10.1016/j.cell.2010.03.031 PMID: 20403328

43. Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol. 2011; 12: R110. https://doi.org/10.1186/gb-2011-12-11-r110 PMID: 22050731

44. Tuller T, Zur H. Multiple roles of the coding sequence 5' end in gene expression regulation. Nucleic Acids Res. 2015; 43: 13–28. https://doi.org/10.1093/nar/gku1313 PMID: 25505165

45. Hardesty B, Tsalkova T, Kramer G. Co-translational folding. Curr Opin Struct Biol. 1999; 9: 111–4. https://doi.org/10.1016/S0959-440X(99)80014-1 PMID: 10047581

46. Nissley DA, O'Brien EP. Timing is everything: unifying codon translation rates and nascent proteome behavior. J Am Chem Soc. 2014; 136: 17892–8. https://doi.org/10.1021/ja510082j PMID: 25486504

47. O'Brien EP, Ciryam P, Vendruscolo M, Dobson CM. Understanding the influence of codon translation rates on cotranslational protein folding. Acc Chem Res. 2014; 47: 1536–44. https://doi.org/10.1021/ar5000117 PMID: 24784899

48. Yu C-H, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, et al. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. Mol Cell. 2015; 59: 744–54. https://doi.org/10.1016/j.molcel.2015.07.018 PMID: 26321254

49. Buhr F, Jha S, Thommen M, Mittelstaet J, Kutz F, Schwalbe H, et al. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. Mol Cell. Elsevier Inc.; 2016; 61: 341–351. https://doi.org/10.1016/j.molcel.2016.01.008 PMID: 26849192

50. Komar AA. The Yin and Yang of codon usage. Hum Mol Genet. 2016; 25: R77–R85. https://doi.org/10.1093/hmg/ddw207 PMID: 27354349

51. Clarke IV TF, Clark PL. Rare codons cluster. PLoS One. 2008;3. https://doi.org/10.1371/journal.pone.0003412 PMID: 18923675

52. Chaney JL, Clark PL. Roles for Synonymous Codon Usage in Protein Biogenesis. Annu Rev Biophys. 2015; 44: 143–166. https://doi.org/10.1146/annurev-biophys-060414-034333 PMID: 25747594

53. Chaney JL, Steele A, Carmichael R, Rodriguez A, Specht AT, Ngo K, et al. Widespread position-specific conservation of synonymous rare codons within coding sequences. PLoS Comput Biol. 2017; 13: 1–19. https://doi.org/10.1371/journal.pcbi.1005531 PMID: 28475588

54. Jacobson GN, Clark PL. Quality over quantity: Optimizing co-translational protein folding with non-'optimal' synonymous codons. Curr Opin Struct Biol. Elsevier Ltd; 2016; 38: 102–110. https://doi.org/10.1016/j.sbi.2016.06.002 PMID: 27318814

55. O'Brien EP, Vendruscolo M, Dobson CM. Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates. Nat Commun. 2014; 5: 2988. https://doi.org/10.1038/ncomms3988 PMID: 24394622

56. Zhang G, Hubalewska M, Ignatova Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat Struct Mol Biol. 2009; 16: 274–80. https://doi.org/10.1038/nsmb.1554 PMID: 19198590

57. Komar AA. A pause for thought along the co-translational folding pathway. Trends Biochem Sci. 2009; 34: 16–24. https://doi.org/10.1016/j.tibs.2008.10.002 PMID: 18996013

58. Thanaraj TA, Argos P. Ribosome-mediated translational pause and protein domain organization. Protein Sci. 1996; 5: 1594–612. https://doi.org/10.1002/pro.5560050814 PMID: 8844849

59. Lam SD, Dawson NL, Das S, Sillitoe I, Ashford P, Lee D, et al. Gene3D: expanding the utility of domain assignments. Nucleic Acids Res. 2016; 44: D404–9. https://doi.org/10.1093/nar/gkv1231 PMID: 26578585

60. Purvis IJ, Bettany AJE, Santiago TC, Coggins JR, Duncan K, Eason R, et al. The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. J Mol Biol. 1987; 193: 413–7. https://doi.org/10.1016/0022-2836(87)90230-0 PMID: 3298659

61. Komar AA, Jaenicke R. Kinetics of translation of gamma B crystallin and its circularly permutated variant in an in vitro cell-free system: possible relations to codon distribution and protein folding. FEBS Lett. 1995; 376: 195–8. https://doi.org/10.1016/0014-5793(95)01275-0 PMID: 7498540

62. Rapoport TA. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. Nature. 2007; 450: 663–9. https://doi.org/10.1038/nature06384 PMID: 18046402

63. Cymer F, von Heijne G, White SH. Mechanisms of integral membrane protein insertion and folding. J Mol Biol. 2015; 427: 999–1022. https://doi.org/10.1016/j.jmb.2014.09.014 PMID: 25277655

64. Fluman N, Navon S, Bibi E, Pilpel Y. mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. Elife. 2014; 3: e03440. https://doi.org/10.7554/eLife.03440 PMID: 25135940

65. Képès F. The "+70 pause": hypothesis of a translational control of membrane protein assembly. J Mol Biol. 1996; 262: 77–86. https://doi.org/10.1006/jmbi.1996.0500 PMID: 8831781

66. Dessen P, Képès F. The PAUSE software for analysis of translational control over protein targeting: application to E. nidulans membrane proteins. Gene. 2000; 244: 89–96. https://doi.org/10.1016/S0378-1119(00)00002-0 PMID: 10689191

67. Nørholm MHH, Light S, Virkki MTI, Elofsson A, von Heijne G, Daley DO. Manipulating the genetic code for membrane protein production: what have we learnt so far? Biochim Biophys Acta. 2012; 1818: 1091–6. https://doi.org/10.1016/j.bbamem.2011.08.018 PMID: 21884679

68. Morgunov AS, Babu MM. Optimizing membrane-protein biogenesis through nonoptimal-codon usage. Nat Struct Mol Biol. 2014; 21: 1023–5. https://doi.org/10.1038/nsmb.2926 PMID: 25469841

69. Pechmann S, Chartron JW, Frydman J. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. Nat Struct Mol Biol. 2014; 21: 1100–5. https://doi.org/10.1038/nsmb.2919 PMID: 25420103

70. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43: D204–12. https://doi.org/10.1093/nar/gku989 PMID: 25348405

71. Bornemann T, Jöckel J, Rodnina M V, Wintermeyer W. Signal sequence-independent membrane targeting of ribosomes containing short nascent peptides within the exit tunnel. Nat Struct Mol Biol. 2008; 15: 494–9. https://doi.org/10.1038/nsmb.1402 PMID: 18391966

72. Tu L, Khanna P, Deutsch C. Transmembrane segments form tertiary hairpins in the folding vestibule of the ribosome. J Mol Biol. 2014; 426: 185–98. https://doi.org/10.1016/j.jmb.2013.09.013 PMID: 24055377

73. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol. 1981; 151: 389–409. Available: http://www.ncbi.nlm.nih.gov/pubmed/6175758 PMID: 6175758

74. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004; 32: 5036–44. https://doi.org/10.1093/nar/gkh834 PMID: 15448185

75. Wilson DN, Arenz S, Beckmann R. Translation regulation via nascent polypeptide-mediated ribosome stalling. Curr Opin Struct Biol. 2016; 37: 123–33. https://doi.org/10.1016/j.sbi.2016.01.008 PMID: 26859868

76. Sabi R, Tuller T. Computational analysis of nascent peptides that induce ribosome stalling and their proteomic distribution in Saccharomyces cerevisiae. RNA. 2017; 23: 983–994. https://doi.org/10.1261/rna.059188.116 PMID: 28363900

77. Gorochowski TE, Ignatova Z, Bovenberg RAL, Roubos JA. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. Nucleic Acids Res. 2015; 43: 3022–32. https://doi.org/10.1093/nar/gkv199 PMID: 25765653

78. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet. 2011; 12: 683–91. https://doi.org/10.1038/nrg3051 PMID: 21878961

79. Tetsch L, Koller C, Haneburger I, Jung K. The membrane-integrated transcriptional activator CadC of Escherichia coli senses lysine indirectly via the interaction with the lysine permease LysP. Mol Microbiol. 2008; 67: 570–83. https://doi.org/10.1111/j.1365-2958.2007.06070.x PMID: 18086202

80. Marman HE, Mey AR, Payne SM. Elongation factor P and modifying enzyme PoxA are necessary for virulence of Shigella flexneri. Infect Immun. 2014; 82: 3612–21. https://doi.org/10.1128/IAI.01532-13 PMID: 24935977

81. Qi F, Frishman D. Melting temperature highlights functionally important RNA structure and sequence elements in yeast mRNA coding regions. Nucleic Acids Res. 2017; 45: 6109–6118. https://doi.org/10.1093/nar/gkx161 PMID: 28335026

82. Mohammad F, Woolstenhulme CJ, Green R, Buskirk AR. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. Cell Rep. 2016; 14: 686–94. https://doi.org/10.1016/j.celrep.2015.12.073 PMID: 26776510

83. Mason N, Ciufo LF, Brown JD. Elongation arrest is a physiologically important function of signal recognition particle. EMBO J. 2000; 19: 4164–74. https://doi.org/10.1093/emboj/19.15.4164 PMID: 10921896

84. Halic M, Becker T, Pool MR, Spahn CMT, Grassucci RA, Frank J, et al. Structure of the signal recognition particle interacting with the elongation-arrested ribosome. Nature. 2004; 427: 808–14. https://doi.org/10.1038/nature02342 PMID: 14985753

85. Schuller AP, Wu CC-C, Dever TE, Buskirk AR, Green R. eIF5A Functions Globally in Translation Elongation and Termination. Mol Cell. Elsevier Inc.; 2017; 66: 194–205.e5. https://doi.org/10.1016/j.molcel.2017.03.003 PMID: 28392174

86. Pelechano V, Alepuz P. eIF5A facilitates translation termination globally and promotes the elongation of many non polyproline-specific tripeptide sequences. Nucleic Acids Res. 2017; 45: 7326–7338. https://doi.org/10.1093/nar/gkx479 PMID: 28549188

87. Altenhoff AM, Škunca N, Glover N, Train C- M, Sueki A, Piližota I, et al. The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic Acids Res. 2015; 43: D240–9. https://doi.org/10.1093/nar/gku1158 PMID: 25399418

88. Clermont O, Christenson JK, Denamur E, Gordon DM. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ Microbiol Rep. 2013; 5: 58–65. https://doi.org/10.1111/1758-2229.12019 PMID: 23757131

89. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000; 16: 276–7. https://doi.org/10.1016/S0168-9525(00)02024-2 PMID: 10827456

90. Nuel G. Significance Score of Motifs in Biological Sequences. In: Mahdavi M A., editor. Bioinformatics —Trends and Methodologies. InTech; 2011. pp. 173–94. https://doi.org/10.5772/18448

91. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011; 7: 539. https://doi.org/10.1038/msb.2011.75 PMID: 21988835

92. Guindon S, Dufayard J- F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59: 307–21. https://doi.org/10.1093/sysbio/syq010 PMID: 20525638

93. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004; 20: 289–90. https://doi.org/10.1093/bioinformatics/btg412 PMID: 14734327

94. Krylov DM, Wolf YI, Rogozin IB, Koonin E V. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 2003; 13: 2229–35. https://doi.org/10.1101/gr.1589103 PMID: 14525925

95. Borenstein E, Shlomi T, Ruppin E, Sharan R. Gene loss rate: a probabilistic measure for the conservation of eukaryotic genes. Nucleic Acids Res. 2007; 35: e7. https://doi.org/10.1093/nar/gkl792 PMID: 17158152

96. Albalat R, Cañestro C. Evolution by gene loss. Nat Rev Genet. 2016; 17: 379–91. https://doi.org/10.1038/nrg.2016.39 PMID: 27087500

97. Wiśniewski JR, Rakus D. Quantitative analysis of the Escherichia coli proteome. Data Br. 2014; 1: 7–11. https://doi.org/10.1016/j.dib.2014.08.004 PMID: 26217677

98. Wiśniewski JR, Rakus D. Multi-enzyme digestion FASP and the "Total Protein Approach"-based absolute quantification of the Escherichia coli proteome. J Proteomics. 2014; 109: 322–31. https://doi.org/10.1016/j.jprot.2014.07.012 PMID: 25063446

99. Glasner JD, Liss P, Plunkett G, Darling A, Prasad T, Rusch M, et al. ASAP, a systematic annotation package for community analysis of genomes. Nucleic Acids Res. 2003; 31: 147–51. https://doi.org/10.1093/nar/gkg125 PMID: 12519969