J Mol Evol
DOI 10.1007/s00239-010-9408-1

Evolution of Prokaryotic Genes by Shift of Stop Codons

Anna A. Vakhrusheva · Marat D. Kazanov · Andrey A. Mironov · Georgii A. Bazykin

Received: 28 July 2010 / Accepted: 29 October 2010

Abstract De novo origin of coding sequence remains an obscure issue in molecular evolution. One of the possible paths for addition (subtraction) of DNA segments to (from) a gene is stop codon shift. Single nucleotide substitutions can destroy the existing stop codon, leading to uninterrupted translation up to the next stop codon in the gene's reading frame, or create a premature stop codon via a nonsense mutation. Furthermore, short indels-caused frameshifts near gene's end may lead to premature stop codons or to translation past the existing stop codon. Here, we describe the evolution of the length of coding sequence of prokaryotic genes by change of positions of stop codons. We observed cases of addition of regions of 3′UTR to genes due to mutations at the existing stop codon, and cases of subtraction of C-terminal coding segments due to nonsense mutations upstream of the stop codon. Many of the observed stop codon shifts cannot be attributed to sequencing errors or rare deleterious variants segregating within bacterial populations. The additions of regions of 3′UTR tend to occur in those genes in which they are facilitated by nearby downstream in-frame triplets which may serve as new stop codons. Conversely, subtractions of coding sequence often give rise to in-frame stop codons located nearby. The amino acid composition of the added region is significantly biased, compared to the overall amino acid composition of the genes. Our results show that in prokaryotes, shift of stop codon is an underappreciated contributor to functional evolution of gene length.

Keywords Stop codons · Evolution · Prokaryotes · Frameshifts · Tandem stop codons · Stop codon read-through · Translation termination · 3′UTR · Gene length

A. A. Vakhrusheva · A. A. Mironov · G. A. Bazykin
Department of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Vorbyevy Gory 1-73, Moscow, Russia 119992

M. D. Kazanov · A. A. Mironov · G. A. Bazykin (✉)
Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karenty pereulok 19, Moscow, Russia 127994
e-mail: gbazykin@iitp.ru

M. D. Kazanov
Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

A. A. Mironov
State Research Institute for Genetics and Selection of Industrial Microorganisms "GosNIIGenetika", 1st Dorozhny proezd 1, Moscow, Russia 117545

Introduction

Although most new genes originate by duplication of pre-existing genes (Ohno 1970; Lynch and Conery 2000; Lynch 2007; Vitreschak et al. 2008), new protein-coding genes have also evolved out of non-coding sequences (Cai et al. 2008; Zhou et al. 2008; Li et al. 2010). However, such cases are rare, because immediate functionality of a long previously non-functional sequence is rarely easily achievable, and translation of a long segment of such sequence is nearly certain to be deleterious.

Conversely, recruitment of short regions of DNA for fulfillment of various functions is wide-spread. This is probably because shorter novel functional regions are more likely to be immediately adaptive, or at least not too deleterious. For example, non-functional DNA readily evolves into short regulatory sequences (Jordan et al. 2003; Silva et al. 2003; Stephen et al. 2008), including transcription factor binding sites (Dermitzakis et al. 2003; Mustonen and

Published online: 17 November 2010

Lassig 2005; Moses et al. 2006; Doniger and Fay 2007; Rodionov 2007) and splicing sites in eukaryotes (Nurtdinov et al. 2007).

Similarly, new segments of previously non-coding DNA may be added to coding sequence. One way in which new sequences can arise in eukaryotes is exonization of introns or intergenic regions (Kreahling and Graveley 2004; Krull et al. 2005; Piriyapongsa et al. 2007a, b; Nurtdinov et al. 2007). Resulting new segments of coding DNA tend to be short (Kondrashov and Koonin 2003; Kurmangaliyev and Gelfand 2008) and are often initially spliced at low frequencies (Artamonova and Gelfand 2007), which may reduce the selective pressure against them. New protein segments can also arise from insertions that get fixed in coding DNA. Fixation of such insertions is facilitated when they are short and have length in multiple of three (Kondrashov and Koonin 2003), and therefore do not distort the reading frame. Non-compensated frameshifts not in multiple of three can also give rise to novel amino acid sequences (Raes and Van de Peer 2005; Kramer et al. 2006; Okamura et al. 2006; Wernegreen et al. 2009; Frenkel and Korotkov 2009).

Another conceivable mechanism for addition of novel segments of DNA into coding sequence is shift of a start (Wilder et al. 2009; Bazykin and Kochetov 2010) or a stop codon. There are at least two reasons to expect the immediate fitness effect of such additions to be modest. First, since the average length of an open reading frame up to the nearest in-frame stop codon under uniform nucleotide composition is only $\sim$20 amino acids, such events can be expected to lengthen the DNA only slightly, and the effect on the function of the resulting protein will usually be minor. Second, negative selection against terminal segments of proteins is usually reduced, compared to the rest of the protein (Shabalina et al. 2004, Ridout et al. 2010).

Weakness of selection against additions implies that they may spread through population by genetic drift. Still, an addition that reaches a substantial frequency in population may be slightly deleterious, neutral, or immediately advantageous, with most additions probably belonging to the first two categories.

If the added sequence is not immediately optimal, some of the subsequent substitutions in the novel coding segments may increase the fitness conferred by the amended protein. Such substitutions can then be picked up by positive selection, which will shape the added sequence.

Addition of novel segments of DNA, as well as subtraction of DNA, by shift of stop codon have recently been described at moderate evolutionary distances in yeast and in mammals (Giacomelli et al. 2007). Here, we study an analogous process in prokaryotic evolution. In prokaryotes, addition of coding DNA may be further facilitated by frequent use of tandem stop codons as backup translation terminators (Nichols 1970; Major et al. 2002). We compare sets of homologous genes in closely related prokaryotic genomes, and find that shifts of stop codons occur even at short evolutionary distances, e.g. between different bacterial strains.

Results

We aligned 7,088 families of homologous genes together with their 3′UTRs from 623 complete prokaryotic genomes, and analyzed all cases in which individual genes of a family had different positions of stop codons in the alignments. We concluded that the stop codon has shifted, and that a segment of coding DNA has been added to (or subtracted from) the gene, if the coding nucleotide sequence immediately upstream of the stop codon of a gene was unambiguously aligned to the non-coding sequence immediately downstream of the stop codon of its homolog.

A total of 6,814,889 pairs of subclusters (see "Methods" section) with different positions of stop codons were found in the analyzed families; this number includes multiple pairs originating from the same family. Of these, only 232 (0.003%) pairs, coming from 205 different families, passed our stringent alignment quality filtering criteria. The vast majority (229, 98.7%) of differences were observed in alignments of orthologs, while the three remaining cases included pairs of paralogous genes from the same organism. In 48 (20.7%) of the filtered pairs of genes with different positions of stop codons, each position was observed in more than one of the homologous genes of this family (hereafter referred to as "supported cases"; Table 1).

We asked what evolutionary events lead to the observed differences in the positions of the stop codons (Table 1). Most of the shifts were due to point mutations either in the stop codons or that gave rise to stop codons. About one-quarter of the cases were caused by indels in coding regions of length not in multiple of 3, leading to frameshifts and use of out-of-frame stop codons. In such cases, the frameshift-causing indel usually occurred close to the original stop codon (median distance: 31 nucleotides for supported cases, 14 nucleotides for unsupported cases), giving rise to a relatively short region of protein translated out-of-frame. In a few remaining cases, a mutation affecting the stop codon and a frameshift were observed simultaneously. The fraction of supported cases was similar for shifts caused by point mutations and indels (χ^2 test, two-tailed $p = 0.45$).

Distribution of the lengths of the region between the two stop codon positions (regions added to or subtracted from the CDS) is shown in Fig. 1. While most of the regions were short, some regions spanned tens of nucleotides. There was no significant difference in the distribution of the lengths of the added/subtracted regions between the

Table 1 Observed events of evolution of coding sequence length by stop codon shift

	Point mutation		Indel		Both	
	Supported	Unsupported	Supported	Unsupported	Supported	Unsupported
Addition	9	23	4	10	0	0
Subtraction	7	22	1	7	1	1
Uncategorized	21	87	4	28	1	6

"Supported" are cases when each position of the stop codon was observed in multiple sequences (see text)

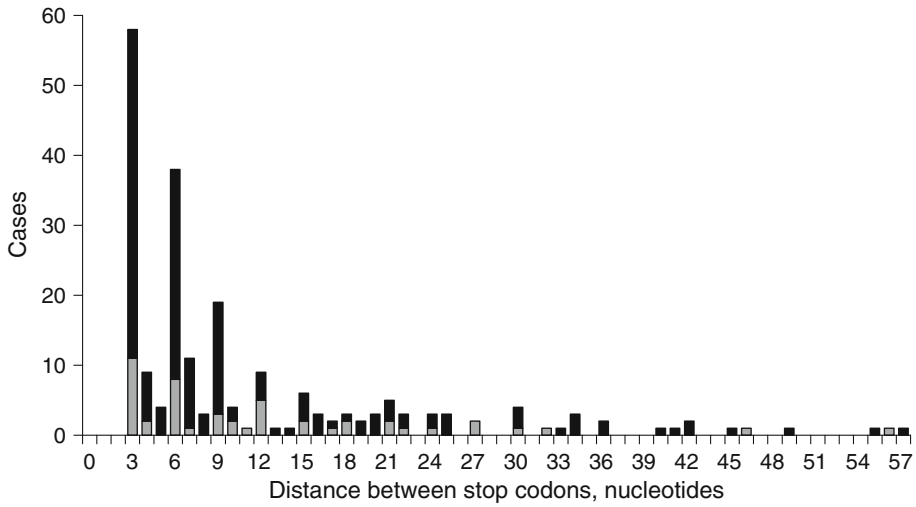

Fig. 1 Shifts of stop codon within families of prokaryotic genes in which each position of stop codon was observed in >1 sequence (*grey* supported) and all other cases (*black* unsupported), for each distance between the stop codons

supported and unsupported sequences (Mann–Whitney U test, $p = 0.19$).

The shifted codons did not appear to be particularly prone to stop codon read-through. The weak UAGC and UAGG contexts of the stop codon (Bertram et al. 2001) were responsible for only 8.4% of the 5′ stop codons of the shifted genes, and were no more frequent in the shifted stop codons than in the rest of the genes (χ^2 test, $p = 0.61$).

In order to distinguish additions from subtractions of C-terminal coding segments, we manually analyzed the phylogenetic distribution of each stop codon position in each of the 232 cases of shift of stop codon. Whenever the phylogeny of the corresponding species was known, we attempted to polarize the corresponding mutations using maximum parsimony. The phylogenetic history of the position of the stop codon could be established unambiguously in 45.8% of supported cases, and in 34.2% of unsupported cases of stop codon shift (Table 1). In approximately half of these cases, the 5′ position of the stop codon was ancestral, implying addition of a region of 3′UTR to the coding sequence (Fig. 2). In the remaining half, the 3′ position of the stop codon was ancestral, implying subtraction of the C-terminal coding segment (Fig. 3). The additions and subtractions were equally frequent (χ^2 test, supported: $p = 0.55$, unsupported: $p = 0.79$).

The median lengths of the added (subtracted) regions was 12.0 (4.0) nucleotides among the supported cases, and 7.0 (6.0) nucleotides among the unsupported cases, but the difference between the lengths of the added and subtracted regions was insignificant (Mann–Whitney U test, supported: $p = 0.08$, unsupported: $p = 0.90$).

In all 13 supported cases of addition, and in 27 (81.8%) of the unsupported cases of addition, the terminating triplet which became the new stop codon during addition predated the mutation disrupting the ancestral stop codon (Fig. 2). In the remaining six (18.2%) of the unsupported cases of addition, the new 3′ stop codon originated de novo from a non-terminating triplet through a point mutation on the same phylogenetic branch as the addition event.

Additions usually occurred in cases when a nearby downstream stop codon allowed it. Indeed, the median distance between the stop codon and the following downstream in-frame TAA, TAG, or TGA triplet for all considered genes was 33.0 nucleotides—only slightly less than the 43.3 nucleotides expected under uniform nucleotide composition. This is despite the large number of genes with tandem stop codons, i.e., those in which the stop codon was immediately followed by another terminating triplet (Fig. 4). In contrast, the median distance to the nearest in-frame TAA, TAG, or TGA triplet for the supported parsimony-reconstructed ancestral genes that subsequently

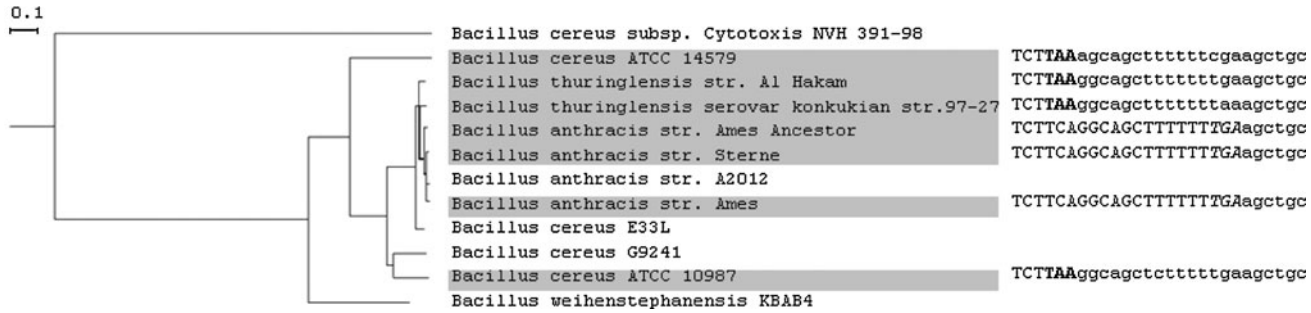

Fig. 2 Example of an addition of a segment of DNA to the coding region in *modB* gene of *Bacillus* genus. *Left* the phylogeny of the considered species in the ATGC database. The species for which the sequence of the gene was present in the alignment are highlighted in *grey*. The edge of the phylogeny at which the addition event happened is highlighted. *Right* nucleotide alignment of the C′-termini of the corresponding genes (*uppercase*) and the upstream region of the 3′UTR (*lowercase*). The ancestral stop codon is in *bold*; the derived stop codon is in *italic*. In this case, the terminating triplet which became the new stop codon during addition predated the mutation disrupting the ancestral stop codon, since the former existed in two outgroup sequences: *B. thuringlensis* str. Al Hakam and *B. cereus* ATCC 10987

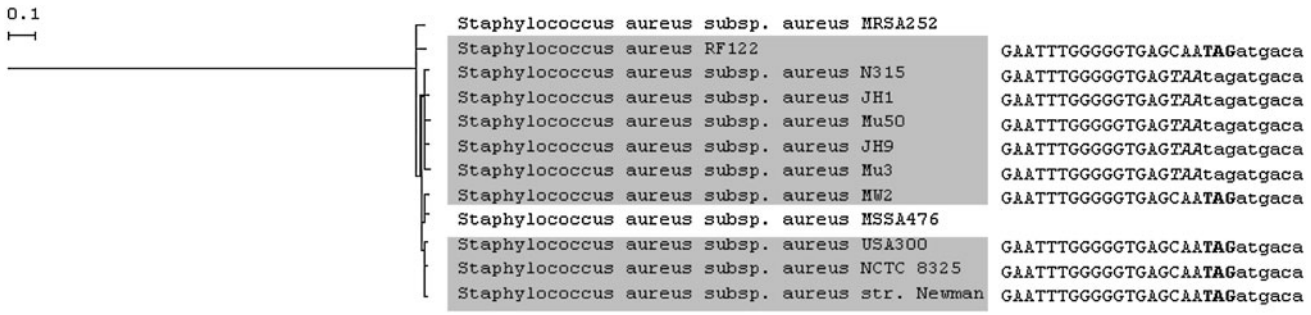

Fig. 3 Example of a subtraction of a segment of DNA from the coding region in *rexB* gene of *Staphylococcus* genus. *Left* the phylogeny of the considered species in the ATGC database. The species for which the sequence of the gene was present in the alignment are highlighted in *grey*. The edge of the phylogeny at which the subtraction event happened is highlighted. *Right* nucleotide alignment of the C′-termini of the corresponding genes (*uppercase*) and the upstream region of the 3′UTR (*lowercase*). The ancestral stop codon is in *bold*; the derived stop codon is in *italic*

experienced a point mutation in the stop codon was only 9.0 nucleotides, i.e. ∼3.5 times lower (Fig. 4a). Similarly, the median distance to the nearest out-of-frame TAA, TAG, or TGA triplet was much lower in the genes with frameshift-caused additions (7.5 for frame+1 and 6.0 for frame+2) than in the rest of the genes (30.0 for frame+1 and 27.0 for frame+2) (Fig. 4b, c).

Since the added regions originate from previously non-coding DNA, we hypothesized that their amino acid composition will be biased, compared to the amino acid composition of the coding regions of the genes as a whole. Indeed, the amino acid composition of the captured regions was significantly different from that of our entire dataset (i.e., all the genes in all the considered families) (χ^2 test, $p < 10^{-6}$). It was also significantly different from the C-terminal 30 amino acids in genes in our dataset (χ^2 test, $p = <10^{-6}$), indicating that the difference in composition is not due to biases inherent to the C-terminus of the protein. Cluster analysis performed on the Euclidian distances between the amino acid composition vectors (Echols et al. 2002) showed that the composition of the captured regions most closely resembled the composition of the translated UTR (Fig. 5).

In most of the observed cases of subtraction, the new stop codon originated by a nonsense point mutation (Table 1; Fig. 3). In two out of seven (four out of 22) such supported (unsupported) cases, the ancestral stop codon mutated to a non-terminating triplet on the same phylogenetic branch as the subtraction event. In the remaining five (18) cases, it remained intact, so that a pair of nearby terminating triplets was established.

Most of the observed stop codon shift events occurred between closely related species. In 12 out of 13 (92.3%) supported additions (32 out of 33 unsupported additions, 96.9%), and in all nine supported subtractions (27 out of 30 unsupported subtractions, 90.0%), the derived long (short) form was nested within a single genus. The close relatedness of the species with different positions of the stop codon impeded the study of subsequent evolution of novel added C-terminal regions of proteins. In nine out of 13 supported additions, and in 19 out of 33 unsupported additions, the added regions were identical between all species supporting the "long" form.

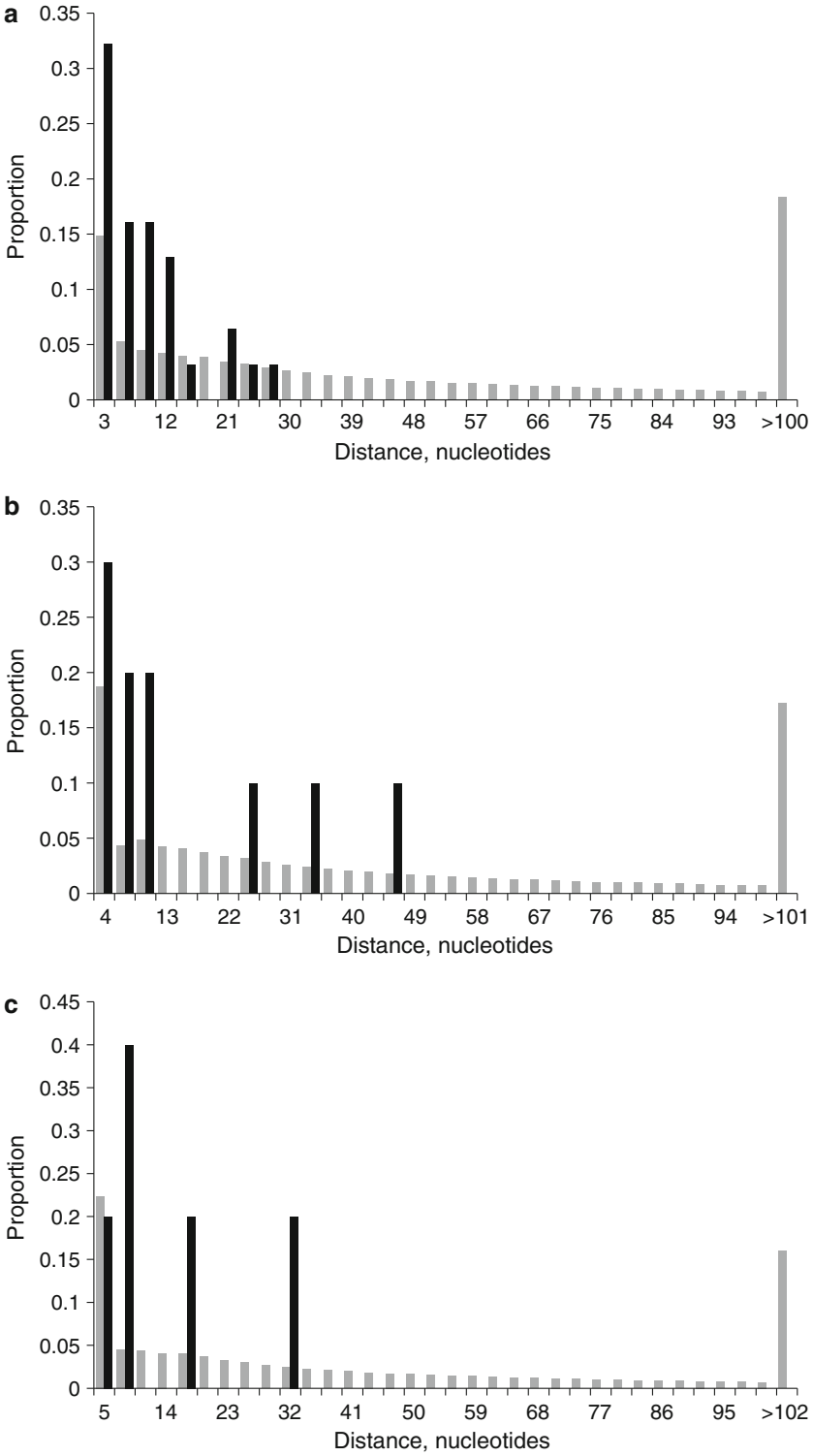

Fig. 4 *Grey* distribution of distances between the stop codon and the next downstream TAA, TGA, or TAG triplet located in frame (**a**), in frame+1 (**b**), or in frame+2 (**c**), in all considered genes. *Black* distribution of lengths of observed cases of addition of 3′UTR segments due to single-nucleotide substitutions in stop codons (**a**), or due to indels leading to translation of the subsequent segment in frame+1 (**b**) or +2 (**c**). Distance 3 corresponds to tandem stop codons, e.g., TAATGA

Discussion

Mismatches in positions of stop codons in homologous genes of different species could be due to annotation or sequencing errors, or could be traces of actual evolution of the C-termini of the genes. However, in-frame stop codons unambiguously terminate translation, so their annotation is rarely uncertain. Although stop codon read-through has been described (Radloff and Kaesberg 1973; Engelberg-Kulka et al. 1977; Rocha et al. 1999; Bertram et al. 2001), it is infrequent, and most stop codons terminate translation very efficiently. As for sequencing errors, they are unlikely

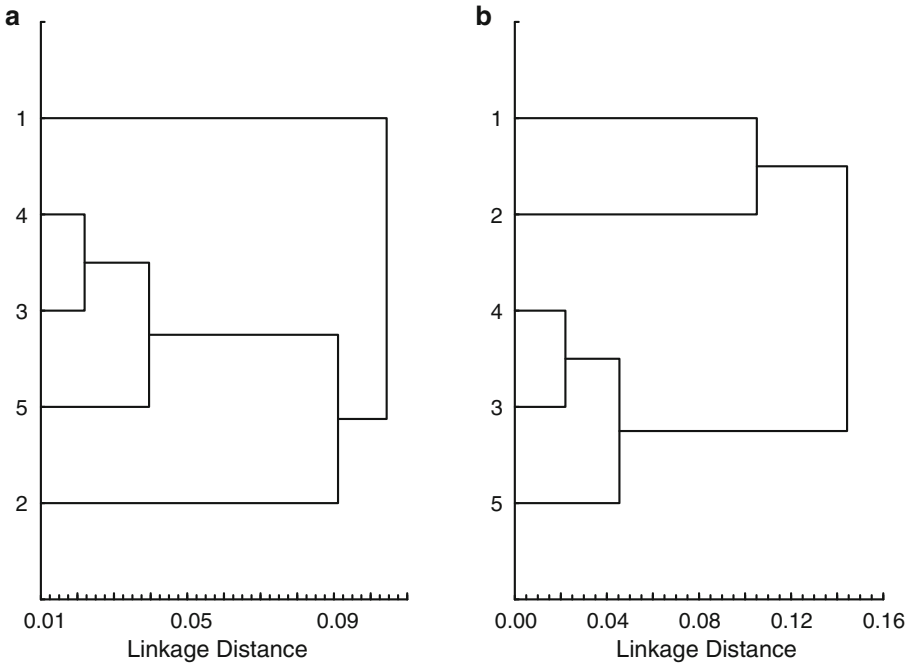

Fig. 5 Results of cluster analysis performed on the Euclidian distances between the amino acid composition vectors using UPGMA (**a**) and Ward's method (**b**). The amino acid composition of the captured regions (*1*) resembles the composition of the proximal 100 amino acids translated from the 3′UTR of the genes in our dataset (*2*) more than that of the 30 C-terminal amino acids in the genes in our dataset (*3*), all genes in our dataset (*4*), or the genes in 15 taxa from three domains of life (*5*; Jordan et al. 2005)

to coincide between organisms. Occurrence of each stop codon position in multiple homologs in a substantial fraction of cases (Table 1) suggests that the change of its position is not attributable to sequencing artifacts, and similarity of the frequencies of supported and unsupported cases among stop codon shifts associated with different mutational mechanisms (Table 1), as well as the similarity of the lengths of the resulting additions and subtractions (Fig. 1), all suggest that such artifacts contribute only little, if anything, to the observed cases. Frequent observation of both stop codon positions in multiple species or strains also means that the shifted stop codons cannot be too deleterious, since strongly deleterious mutations are unlikely to reach high frequencies in bacterial populations and get fixed in interspecies evolution.

Therefore, our results indicate that the position of a stop codon is evolutionarily labile in prokaryotes. This result parallels that obtained for eukaryotes (Giacomelli et al. 2007). The stop codon shifts did not necessarily have a major effect on the produced protein if the 5′ stop codon position could be read through, since in this case the long form would have been produced both before and after the shift. However, such cases were rare, implying that stop codon shifts usually corresponded to changes in gene lengths.

The vast majority of the observed addition and subtraction events happened relatively recently—often between different strains of a single bacterial species. This may partially reflect an ascertainment bias in the choice of the analyzed genomes: the genomes in our dataset usually were either phylogenetically remote (and did not pass our filtering criteria), or very close, often belonging to different strains of the same well-studied species, e.g., *E. coli*. The occurrence of multiple cases of stop codon shift at such close phylogenetic distances implies that this phenomenon is also widespread in evolution on large phylogenetic distances. In fact, many plausible cases of change of the stop codon position between phylogenetically remote species were observed in our dataset, but did not pass our stringent filtering criteria. Furthermore, multiple shift events were observed in a substantial fraction of the families in which a single event was observed, suggesting that some genes may be especially prone to variability in stop codon position.

A point mutation of a stop codon is a simple evolutionary path to obtaining a novel coding sequence. Similarly to novel exons in eukaryotes, novel terminal-added regions can serve as the raw material for natural selection to act upon. We observed that such gained amino acid sequences have a characteristic amino acid composition, distinct from that of the rest of the genome, and resembling that of the translated 3′UTR. It would be interesting to study how such novel protein segments evolve subsequently to their origin. Unfortunately, this would require annotation of ancient capture events, which is hard, especially in prokaryotes, because alignment of non-coding regions becomes problematic with increasing phylogenetic distance. Denser sequencing of bacterial phylogenies may help alleviate this problem and allow the study of subsequent evolution of the gained regions.

Usually, the would-be stop codon exists in 3′UTR by the time a new sequence is added to the C-terminus of a gene. Tandem (Fig. 3) and other nearby in-frame stop codons are conserved and may have a role in prevention of translation

read-through (Nichols 1970; Major et al. 2002; Liang et al. 2005). Our results suggest that read-through prevention notwithstanding, such nearby stop codons may serve as preadaptations that facilitate elongation of the coding sequence. Furthermore, newly acquired stop codons from previous 3′UTR may preferentially occur nearby (Fig. 4) simply because relatively short ORF additions have a higher likelihood of being advantageous (or at least selectively neutral) than long ORF additions.

Conversely, coding regions may be subtracted from a gene due to nonsense mutations. Such mutations can get fixed if selection against them is not too strong; the latter is likely to be the case when the nonsense mutation occurs not far from the original stop codon. Since the original stop codon remains intact in the process, the fixation of nonsense mutations in C-terminal region of a protein may have a side effect of creating nearby terminating triplets which, in turn, may serve as backup translation terminators.

The origin and maintenance of tandem stop codons in prokaryotes is enigmatic, given their relatively low efficiency in translation read-through prevention (Major et al. 2002). Our results suggest that there may be another reason for the overrepresentation of tandem stop codons: they may be the footprints of past subtraction events, in addition to, or rather than, a selected functional trait.

Methods

Homologous prokaryotic genes were obtained from the COG database (Tatusov et al. 2001) and extended onto 623 complete prokaryotic genomes (Kazanov 2008). Briefly, the COG clusters extension procedure was as follows. All sequenced prokaryotic genomes that are not included in the COG database were downloaded from NCBI. Each protein-coding sequence from those genomes was then compared to the COG database proteins using BLASTP (Altschul et al. 1990). The BLASTP output was filtered according to the following criteria: percentage of sequence identity more than 20%; expect value less than 10^{-5}; alignment region covers at least two-thirds of the COG protein length. Next, the filtered BLASTP hits list was sorted by percentage of sequence identity in descending order, and the sequence in consideration was assigned to the topmost COG cluster. The described procedure was tested by a tenfold cross-validation technique as follows. All 144,320 COG database proteins were randomly divided into ten groups, and the above algorithm was applied to each group, while treating the other nine groups as the COG sequence database. This test demonstrated the 97.5% accuracy of the developed COG assignment procedure.

In each family, the protein-coding sequences of each gene, together with 300 nucleotides downstream of the stop codon, were extracted from the corresponding prokaryotic genomes. Nucleotide alignments of all genes and downstream regions in each family were then built with MUSCLE (Edgar 2004).

All sequences within each multiple alignment were grouped into subclusters based on the alignment positions of the stop codons. Each pair of subclusters with different positions of stop codons was then analyzed separately. To distinguish the cases of actual shift of the position of the stops from regions with problematic alignments, a conservative set of criteria was applied to each pair of sequences from different subclusters, i.e., to each putative stop codon shift. After some experimenting, the following filtering criteria were chosen to exclude the problematic alignment regions.

We required the nucleotide identity of the whole CDS between the two sequences to be more than 35%, and the fraction of gaps in alignment between the two positions of stop codons in the first and in the second sequence to be no more than 30%. Further, we required 10 matches in 15 nucleotides before the 5′ stop codon position, and eight matches in 15 nucleotides after the 3′ stop codon position. Among pairs of subclusters, only such were chosen in which at least one pair of species (one from each subcluster) met the above criteria. All the resulting cases of stop codon shift were examined manually to ensure that the above criteria filtered out the problematic cases.

For each position of the stop codon, we asked whether there were other sequences with a stop codon at the same position, i.e. whether the corresponding subcluster included more than one sequence. We considered a position of the stop "supported" if, in addition to the sequence chosen at the previous step, there were other sequences in the subcluster with at least five matches among 10 nucleotides before and five matches among 10 nucleotides after the position of the stop.

Whenever possible, each case of change of position of stop codon was categorized as "addition" or "subtraction" of non-coding DNA. For this purpose, we manually mapped the species with each position of the stop codon onto the prokaryotic phylogeny, and inferred the direction of the change using maximum parsimony. This polarization allowed us to distinguish subtractions of C-terminal coding segments from additions of regions of 3′UTR into a gene. In doing so, we assumed that the gene trees for the considered genes coincide with the species tree, i.e. did not account for the possibility of lateral gene transfer. However, since most cases of stop codon shift involved very closely related species or strains, lateral transfers should not substantially bias our results, and in general maximum parsimony approach can be expected to be accurate. The phylogeny of the prokaryotic species was taken from ATGC database (http://atgc.lbl.gov/atgc/; Novichkov et al. 2009).

The expectation for the median length M of amino acid sequence before the first in-frame stop codon under uniform nucleotide composition is the median of the geometric distribution with mean 64/3 and equals:

$$M = \frac{-\log 2}{\log(1-P)} \sim 14.44,$$

where $p = 3/64$ is the probability of a stop codon at each triplet. The expected median length of the corresponding nucleotide sequence is 43.32.

The Euclidian distance D between the amino acid composition vectors was calculated as

$$D = \sqrt{\sum_{i=1}^{N} [F_i(A) - F_i(B)]^2}$$

(Echols et al. 2002), where $N = 20$ is the number of amino acids, and $F_i(A)$ is the frequency of amino acid i under condition A. The conditions were then clustered using the unweighted pair-group averaging method (UPGMA) and Ward's method (Ward 1963) based on the Euclidian distances between them.

Acknowledgments This work was supported by the grants from the Russian Foundation for Basic Research [08-04-01394-a], the Russian Ministry of Science and Education grant "Phylogenetic analysis of complex selection in molecular evolution" and contract P916, and the "Molecular and Cellular Biology" Program of the Russian Academy of Sciences.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Artamonova II, Gelfand MS (2007) Comparative genomics and evolution of alternative splicing: the pessimists' science. Chem Rev 107:3407–3430

Bazykin G, Kochetov A (2010) Alternative translation start sites are conserved in eukaryotic genomes. Nucl Acids Res. doi:10.1093/nar/gkq806

Bertram G, Innes S, Minella O, Richardson J, Stansfield I (2001) Endless possibilities: translation termination and stop codon recognition. Microbiology 147:255–269

Cai J, Zhao R, Jiang H, Wang W (2008) De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. Genetics 179:487–496

Dermitzakis ET, Bergman CM, Clark AG (2003) Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites. Mol Biol Evol 20:703–714

Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol 3:e99

Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z, Gerstein M (2002) Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. Nucl Acids Res 30:2515–2523

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 32:1792–1797

Engelberg-Kulka H, Dekel L, Israeli-Reches M (1977) Streptomycin-resistant *Escherichia coli* mutant temperature sensitive for the production of Qbeta-infective particles. J Virol 21:1–6

Frenkel FE and Korotkov EV (2009) Using triplet periodicity of nucleotide sequences for finding potential reading frame shifts in genes. DNA Res 16:105–114

Giacomelli MG, Hancock AS, Masel J (2007) The conversion of 3′UTR's into coding regions. Mol Biol Evol 24:457–464

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19:68–72

Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S (2005) A universal trend of amino acid gain and loss in protein evolution. Nature 433:633–638

Kazanov MD (2008) Functional classification of genes from complete bacterial genomes, based on Clusters of Orthologous Groups (COG) database. In: Proceedings of informational technologies and systems conference, Moscow, pp 104–109

Kondrashov FA, Koonin EV (2003) Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. Trends Genet 19:115–119

Kramer EM, Su HJ, Wu CC, Hu JM (2006) A simplified explanation for the frameshift mutation that created a novel C-terminal motif in the *APETALA3* gene lineage. BMC Evol Biol 6:30

Kreahling J, Graveley BR (2004) The origins and implications of aluternative splicing. Trends Genet 20:1–4

Krull M, Brosius J, Schmitz J (2005) Alu-SINE exonization: en route to protein-coding function. Mol Biol Evol 22:1702–1711

Kurmangaliyev YZ, Gelfand MS (2008) Computational analysis of splicing errors and mutations in human transcripts. BMC Genomic 9:13

Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, Lu SJ, Li XM, Yu Q, Zheng X, Du Q, Uhl GR, Liu QR, Wei L (2010) A human-specific de novo protein-coding gene associated with human brain functions. PLoS Comput Biol 6:e1000734

Liang H, Cavalcanti AR, Landweber L (2005) Conservation of tandem stop codons in yeasts. Genome Biol 6:R31

Lynch M (2007) The origins of genome architecture. Sinauer Assosiates, Sunderland

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Major LL, Edgar TD, Yee YP, Isaksson LA, Tate WP (2002) Tandem termination signals: myth or reality? FEBS Lett 514:84–89

Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB (2006) Large-scale turnover of functional transcription factor binding sites in *Drosophila*. PLoS Comput Biol 2:e130

Mustonen V, Lassig M (2005) Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. Proc Natl Acad Sci 102:15936–15941

Nichols JL (1970) Nucleotide sequence from the polypeptide chain termination region of the coat protein cistron in bacteriophage R17 RNA. Nature 225:147–151

Novichkov PS, Ratnere I, Wolf YI, Koonin EV, Dubchak I (2009) ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. Nucl Acids Res 37:D448–D454

Nurtdinov RN, Neverov AD, Favorov AV, Mironov AA, Gelfand MS (2007) Conserved and species-specific alternative splicing in mammalian genomes. BMC Evol Biol 7:249

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York

Okamura K, Feuk L, Marquès-Bonetc T, Navarroc A, Scherer SW (2006) Frequent appearance of novel protein-coding sequences by frameshift translation. Genomics 88:690–697

Piriyapongsa J, Polavarapu N, Borodovsky M, McDonald J (2007a) Exonization of the LTR transposable elements in human genome. BMC Genomics 8:291

Piriyapongsa J, Rutledge MT, Patel S, Borodovsky M, Jordan IK (2007b) Evaluating the protein coding potential of exonized transposable element sequences. Biol Direct 2:31

Radloff RJ, Kaesberg P (1973) Electrophoretic and other properties of bacteriophage Q: the effect of a variable number of read-through proteins. J Virol 11:116–128

Raes J, Van de Peer Y (2005) Functional divergence of proteins through frameshift mutations. Trends Genet 21:428–431

Ridout KE, Dixon CJ, Filatov DA (2010) Positive selection differs between protein secondary structure elements in *Drosophila*. Genome Biol Evol 2:166–179

Rocha EP, Danchin A, Viari A (1999) Translation in *Bacillus subtilis:* roles and trends of initiation and termination, insights from a genome analysis. Nucl Acids Res 27:3567–3576

Rodionov DA (2007) Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. Chem Rev 107:3467–3497

Shabalina SA, Ogurtsov AY, Rogozin IB, Koonin EV, Lipman DJ (2004) Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. Nucl Acids Res 32: 1774–1782

Silva JC, Shabalina SA, Harris DG, Spouge JL, Kondrashov AS (2003) Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. Genet Res 82:1–18

Stephen S, Pheasant M, Makunin IV, Mattick JS (2008) Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol Biol Evol 25:402–408

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucl Acids Res 29:22–28

Vitreschak AG, Mironov AA, Lyubetsky VA, Gelfand MS (2008) Comparative genomic analysis of T-box regulatory systems in bacteria. RNA 14:717–735

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Wernegreen JJ, Kauppinen SN, Degnan PH (2009) Slip into something more functional: selection maintains ancient frameshifts in homopolymeric sequences. Mol Biol Evol 27:833–839

Wilder JA, Hewett EK, Gansner ME (2009) Molecular evolution of GYPC: evidence for recent structural innovation and positive selection in humans. Mol Biol Evol 26:2679–2687

Zhou Q, Zhang Z, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W (2008) On the origin of new genes in *Drosophila*. Genome Res 18:1446–1455