Nucleic Acids Research, 1999, Vol. 27, No. 17 3567–3576

Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis

Eduardo P. C. Rocha[1,2,*], Antoine Danchin[2] and Alain Viari[1]

[1]Atelier de BioInformatique, Université Paris VI, 12 Rue Cuvier, 75005 Paris, France and [2]Unité de Regulation de l'Expression Génétique, Institut Pasteur, 28 Rue Dr Roux, 75724 Paris, France

Received April 19, 1999; Revised and Accepted July 16, 1999

ABSTRACT

We analysed the *Bacillus subtilis* protein coding sequences termini, and compared it to other genomes. The analysis focused on signals, compositional biases of nucleotides, oligonucleotides, codons and amino acids and mRNA secondary structure. AUG is the preferred start codon in all genomes, independent of their G+C content, and seems to induce less stable mRNA structures. However, it is not conserved between homologous genes neither is it preferred in highly expressed genes. In *B.subtilis* the ribosome binding site is very strong. We found that downstream boxes do not seem to exist either in *Escherichia coli* or in *B.subtilis*. UAA stop codon usage is correlated with the G+C content and is strongly selected in highly expressed genes. We found less stable mRNA structures at both termini, which we related to mRNA–ribosome and mRNA–release-factor interactions. This pattern seems to impose a peculiar A-rich nucleotide and codon usage bias in these regions. Finally the analysis of all proteins from *B.subtilis* revealed a similar amino acid bias near both termini of proteins consisting of over-representation of hydrophilic residues. This bias near the stop codon is partially release-factor specific.

INTRODUCTION

Translation is a very energy-demanding process which in bacteria consumes most of the metabolic resources (1). Therefore, translation patterns in genes within a particular environmental context are expected to be significantly constrained by criteria of efficient or fast translation (2). In the present analysis of translation patterns in *Bacillus subtilis* we have followed the usual division of the process into three consecutive phases: initiation, elongation and termination.

The initiation phase is regarded as rate limiting in most cases (3,4). The key intermediate of this phase is the formation of the 30S initiation complex, containing the 30S ribosomal subunit, mRNA, fMet-tRNA$_f$, and the three initiation factors (5). The free 3'-end of the 16S rRNA plays a critical role in the initiation of protein synthesis by base-pairing with the complementary ribosome binding site (RBS) upstream of the start codon in the mRNA. In *Escherichia coli* the protein S1 is known to play a crucial role in the attachment of the mRNA to the 16S subunit of the ribosome complex (4). The absence of a homologous protein in a large group of Gram-positive eubacteria, such as *B.subtilis*, appears to magnify the importance of a strong RBS and leads to poor expression of most *E.coli* genes in these organisms (6). The 70S complex is formed after the correct placement of the initiation tRNA with respect to the first codon (7). Among the elements that may play additional roles in the initiation is the level of mRNA structure at the RBS (3), and at the beginning of the reading frame (8), protein–mRNA interactions (7), codon bias at the beginning of genes (9), and putative signals such as the 'downstream box', that supposedly interacts with the 16S subunit at UCAUCUGUCCCACCU (10).

The elongation phase proceeds relatively quickly, though it can be retarded or stalled by the existence of mRNA structures (11), and the use of rare codons (12,13). For a given species' G+C content (14), the codon usage reflects the relative amounts of the different tRNA species (15), and the efficiency of codon–anti-codon interaction (2). The existence of stable mRNA secondary structures within the coding sequence may stall or even abort the translation, possibly implying faster mRNA degradation (16).

The termination depends upon the attachment of a release factor (RF) in the place of a tRNA in the ribosomal complex. RF1 recognises UAA and UAG, and RF2 recognises UAA and UGA. A third factor (RF3) is dispensable, not codon specific, and is known to stimulate the activities of the other two factors (17). If the interaction between the RF and the ribosome is slow, the exposed stop codon can be recognised by a near cognate tRNA giving an elongated protein as a readthrough product (18). Incorrect termination will not only probably lead to a defective protein, but also to an important waste of resources, since it takes place after the entire transcription and translation of the gene. In *E.coli* the efficiency of the UGA codon is context dependent, since the two last amino acids of the protein act co-operatively towards the efficiency of termination (19). This seems to be related with the van der Waals volume of the last amino acid and the hydrophobicity of the penultimate amino acid (19). Moreover, important nucleotide

*To whom correspondence should be addressed at: Atelier de Bioinformatique, Université Paris VI, 12 Rue Cuvier, 75005 Paris, France. Tel: +33 1 44 276536; Fax: +33 1 44 276312; Email: erocha@abi.snv.jussieu.fr

biases have been identified downstream of the stop codon and may reflect sites of contact between the RF and the mRNA (18,20).

In this study, we have analysed each phase in terms of signals, compositional biases and mRNA secondary structures. A 'signal' is a word with one or a few variants (e.g. RBS, start and stop codons). Our main goals were to perform an extensive analysis of translation patterns in the complete genome of *B.subtilis* and to clarify their roles at these different phases. Most studies done so far on this subject have been devoted to only one of the phases and almost exclusively to *E.coli* (9,21–24). Therefore, we have searched to correlate the diverse pieces of information in order to sketch a general integrative picture of this process in *B.subtilis*. When our results suggested disagreement with results published for *E.coli* (or when such results did not exist) we compared the two genomes.

MATERIALS AND METHODS

Sequences

Sequences and annotations of *B.subtilis* (25) and *E.coli* (26) were taken from the Subtilist and Colibri databases (27; http://www.pasteur.fr/Bio/). Information on *B.subtilis* proteins was taken from Swissprot release 34 (http://www.expasy.ch) (28). Data on the remaining complete genomes was taken from the Entrez Genome Browser (http://www.ncbi.nih.gov). To test biological hypotheses we built data sets representing given properties (e.g. genes ending in UAG), which are compared to the complementary data sets. To analyse patterns related to membrane and exported proteins we have built an export set (with genes classified as cell envelope and cellular processes in ref. 25), and a non-export set (genes of central metabolism in ref. 25, excluding lipids, specific pathways and amino acid metabolism). We verified that the genes in the non-export set did not contain any gene with predicted or verified signal peptide in the Swissprot database.

Codon usage

We used the previously published classification of genes using factorial correspondence analysis (FCA) of codon usage bias in three classes. This classification is correlated to gene expression level: moderately expressed (class 1, 3375 genes), highly expressed under exponential growth conditions (class 2, 188 genes) and of likely foreign origin (class 3, 537 genes) (27). This method provides a classification similar to other methods [e.g. RSCU and CAI (29)], with similar caveats. In particular, 84% of ribosomal genes fall in class 2, but 91% of tRNA synthetase genes, which are also highly expressed, fall in class 1. Both groups are predominantly in the leading strand of the chromosome (93% of these genes) hence the difference is not due to strand bias (30). To circumvent this problem we built two additional data sets of ribosomal proteins (56 genes) and tRNA synthetases (25 genes). Whenever a correlation is observed with class 2 genes, suggesting a relation with high degree of expression, it is also checked against these data sets.

mRNA structure

The mRNA secondary structure was computed using the libraries of the Vienna package (31), that computes the energy of the best structure using the method developed by Zuker and Stiegler (32). We have used the default temperature of folding (37ºC) and allowed for G–U pairs. A first analysis was done through the folding of 50 bp sliding windows (step of 10 bp) on the gene, separately considering a start region (–100 bp to +100 bp around the start), a middle region (from 100 bp after the start to –100 before the stop) and a stop region (–100 bp to +100 bp around the stop). This analysis provided a measure of the folding potential, which was compared to the energy computed for random sequences with the same nucleotide or trinucleotide composition. None of these foldings has a meaning on its own, since we have not folded the entire mRNA molecule and we ignore pseudo-knots; however, they provide a measure of the propensity to make structures in the different mRNA regions. To inspect more closely the energies involving a signal (start, stop or RBS), we proceeded as follows: we folded a 50 bp window centred on the signal, then we made *in silico* mutations on the signal and finally we folded the window again. The contrasts correspond to energy differences between alternative signals (e.g. UAA and UAG). Small differences found in these analyses can be relevant (if statistically significant), because the stability of a structure increases exponentially with the energy (33).

Building matrices for redundant signals

We have used an in-house implementation of the expected maximisation algorithm (34), in order to define score matrices for non-aligned sequences. This approach allows for the simultaneous identification of the sites and characterisation of the binding motifs as a position scoring matrix. After the definition of the matrix we scored all sequences using standard Bayesian estimators, taking the best hit on the sequence as the score of the match. Probability of the hit is given by the ratio of the matrix score (product of the probabilities at each position of the matrix) to the sum of the matrix score and the background probability.

Set of homologous genes

We have built a set of 7479 genes from other complete genomes presenting high similarity to 1917 genes of *B.subtilis*. This was done using gapped BlastP (35) on all *B.subtilis* proteins, selecting for each species the best hit provided it had a P-value $< 10^{-10}$. Then we further refined this selection by keeping only pairs of homologues of similar length (<10% variation).

Oligonucleotide bias analysis

Word bias. The significance of under- or over-representation of a strict word was defined by comparison to a Markov chain model of maximal order as explained previously (36,37).

Single outliers. The significance of a nucleotide bias by comparison to an average trend (e.g. U after a stop codon) was evaluated through Student's t-test where the mean and variance were taken from the regular set (i.e. the average distribution of U in intergenic positions).

Contrast between distributions. The significance of the contrast between two sets (e.g. lysine before UAG or before UGA), was assigned through the use of a contingency table. The outliers associated with each variable were taken as significant when their contribution to the χ^2 was larger than 3.8 (significance at 1% level in a χ^2 with 1 d.f.).

	1	2	3	4	5	6	7	8
A	0.52	0.52	0.71	0.05	0.04	0.70	0.06	0.23
C	0.08	0.09	0.04	0.01	0.02	0.02	0.01	0.07
G	0.19	0.22	0.16	0.93	0.94	0.16	0.87	0.58
T	0.21	0.14	0.09	0.01	0.00	0.13	0.06	0.12

Class	AUG	GUG	UUG	Other
1	78%	10%	12%	0%
2	86%	8%	6%	0%
3	75%	7%	16%	2%
all	78%	9%	13%	<1%

Class	UAA	UAG	UGA
1	60%	25%	15%
2	91%	4%	5%
3	62%	17%	21%
all	62%	14%	24%

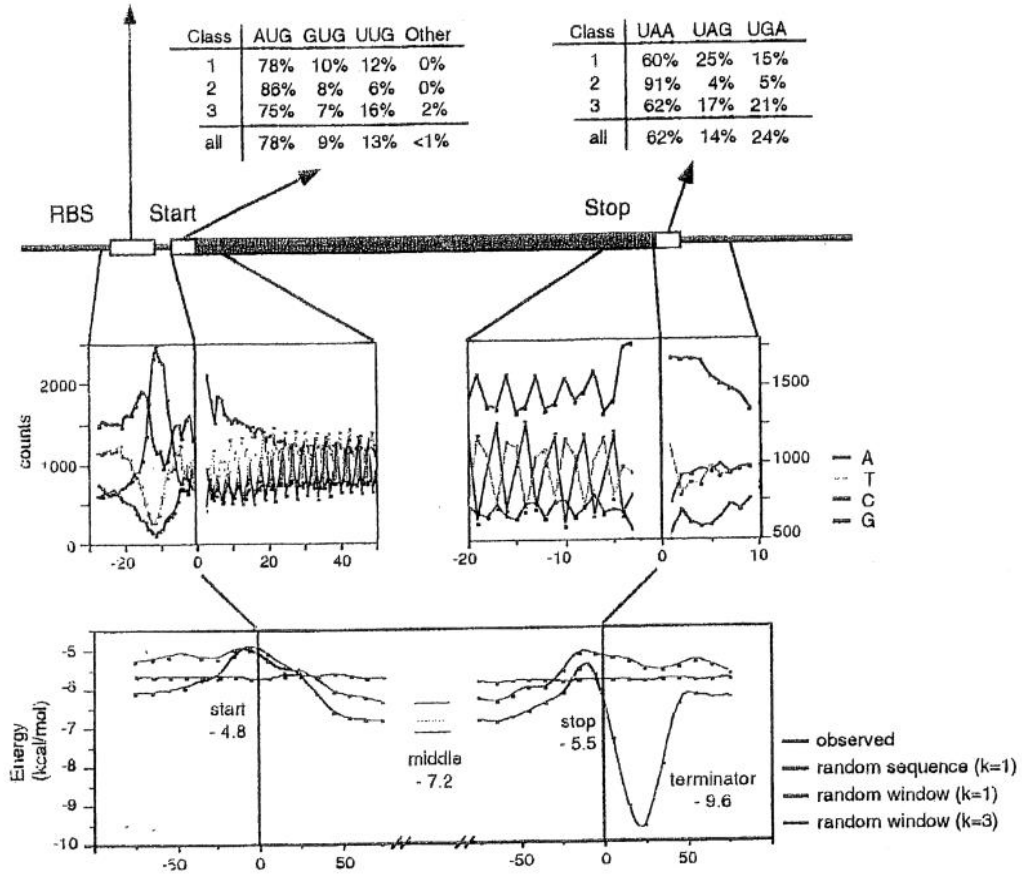

Figure 1. General results of the analysis at the translation level. Consensus matrix of the RBS and statistics on the frequencies of start and stop codons according to the different codon usage classes (top). Distribution of nucleotides near the start and stop of genes (start and stop codons removed for simplification). Curves of the observed mRNA folding energy in sliding windows of 50 bp near the start (from −100 bp to +100 bp), middle and stop codons (from −100 bp to +100 bp), and energy of the corresponding sequences when randomised either completely or in windows keeping the frequency of mono and trinucleotides (bottom).

Correlation analysis

Independence between categorical data (e.g. between the use of start and stop codons), was tested by contingency tables. Correlation between a categorical element and a continuous one (e.g. start codon and *P*-value of the RBS), was performed using successively the Kruskal–Wallis and the Tukey–Kramer test (38). Comparison between continuous variables was performed using multivariate and pairwise correlation analysis (e.g. RBS *P*-value and mRNA structure energy). Relevant outliers were checked through the use of robust association measures and tests, such as the Spearman's rank or the Kendall-τ (38). Analysis of the outliers highlighted elements deviating from the average trend. Unless stated otherwise, statistical significance in all tests was taken at the conservative 1‰ level.

RESULTS AND DISCUSSION

RBS are very strong in *B.subtilis* and alternative downstream boxes do not seem to exist

The RBS matrix obtained using the 30 bp before the start of the *B.subtilis* genes whose products are present in Swissprot is similar to the one obtained using all genes (Fig. 1). About 90% of the RBS are distributed between positions −5 and −11, with mode at −8 (for 23% of the genes; data not shown). This is consistent with experimental results showing an optimal RBS placement at positions 7–9 bp (6). Most RBS are very strong, i.e. close to the consensus sequence AAAGGAGG, and their strength is not correlated with the codon usage class of the gene. If one considers possible translational coupling, and bad start assignment (i.e. there is a good RBS followed by a start

Table 1. Percentage of each start and stop codon in various species and species G+C content

	Start codons (%)				Stop codons (%)			G+C (%)
	AUG	GUG	UUG	other	UAA	UAG	UGA	
Aquifex aeolicus	82	10	7	<1	51	12	37	43
Archaeoglobus fulgidus	76	22	2	0	34	16	51	49
Bacillus subtilis	78	9	13	<1	62	14	24	44
Borrelia burgdorferi	69	9	22	0	63	19	18	29
Chlamydia trachomatis	89	8	4	0	55	30	15	41
Escherichia coli	83	14	3	<1	63	8	29	51
Haemophilus influenzae	89	11	0	0	75	14	11	38
Helicobacter pylori	82	10	8	<1	56	17	27	39
Methanococcus jannaschii	67	15	18	0	78	10	13	31
Methanobacterium thermoautotrophicum	62	22	15	1	30	23	47	50
Mycoplasma genitalium	92	8	0	0	73	27	–	32
Mycoplasma pneumoniae	92	4	3	<1	72	28	–	40
Mycobacterium tuberculosis	61	33	5	1	16	29	55	66
Pyrococcus horikoshii	93	6	0	1	66	18	16	42
Rickettsia prowazekii	83	17	0	0	44	36	20	29
Synechocystis sp.	58	34	8	<1	24	39	37	48
Treponema pallidum	82	10	7	<1	51	12	37	53

codon in phase before or after the annotated start), then only 55 *B.subtilis* genes do not present a good RBS. Many of these may be pseudo-genes, though 20 of them are present in the Swissprot database.

Using positions +4 to +33 of all genes, we failed to characterise a 'downstream box' in a set including all Swissprot genes. Neither did we find it in a set including only Swissprot genes with weak or no RBS. Finally we analysed the *E.coli* tRNA synthetase genes (39), and also failed to uncover significant motifs. Since these were the genes used for the definition of the 'downstream box' we tried a different analysis to verify the significance of the patterns. The best examples of this signal have seven and eight exact hits in a motif of 12 nucleotides (39). We randomised these sequences (1000 experiences for each sequence) respecting the content in nucleotides and checked for the maximal scores we would obtain by chance alone. We observed that 34% of the random sequences had one signal with eight or more matches and 9% had nine matches or more. Therefore, the 'downstream box' patterns are not statistically significant. This confirms results concerning the non-specificity of the toe-printing interactions (40) and the absence of important signals at the start besides the RBS (8).

Though preferred and efficient, AUG is not conserved between homologues nor biased in highly expressed genes

AUG is the predominant start codon, particularly in eubacteria (Table 1). In *B.subtilis* the order of frequencies is AUG > UUG > GUG, whereas in *E.coli* it is AUG > GUG > UUG, which corresponds to the order of degree of expression of these codons in the respective bacteria (6). The relative frequency of start codons on a genome and its G+C content are not significantly correlated (using Kendall-τ and Spearman's rank). Using the set of homologues we computed the choice of the start codon in the *B.subtilis* genes and its homologues (Table 2). We observe that the existence of a start codon in a homologue is independent of its existence in the *B.subtilis* gene. Since the homologous set is made of very different taxa we confirmed these results using only genes homologous to *E.coli* genes. We also checked that this independence holds in the small subset of highly expressed genes (data not shown).

Energy of the mRNA structure near the start codon depends on the start codon, and is significantly higher for AUG starting genes (Table 3). On average the *in silico* mutation of an AUG implies significantly more stable structures (energy 4% lower for UUG and 7% for GUG). Nevertheless, the average energy for these 'mutated' genes is higher than for the average genes effectively starting by UUG and GUG (diagonal elements of Table 3). GUG and UUG starting genes acquire less stable structures when mutated to AUG, but these structures are nevertheless more stable than the ones of true AUG genes. Therefore, it is likely that start codon frequency is partly a result of mRNA structure avoidance.

Several experimental works have demonstrated that the use of AUG increments the degree of expression (6,41,42), we also observe that AUG is related to less stable mRNA structures, and that AUG is systematically preferred in all genomes. This strongly suggests that AUG is positively selected. However, the slight increase of AUG starts in class 2 is not statistically significant ($P > 0.05$) (Fig. 1), and therefore highly expressed genes do not prefer AUG. Moreover the start codon is not conserved among homologous genes, not even among well conserved and highly expressed genes such as the genes for ribosomal proteins. This contradicts the previous statement since it suggests no selective advantage for AUG.

Table 2. Start and stop codon usage in the *B.subtilis* genes (rows) and homologues in other complete genomes (columns)

Bs/Oth	Start codons (%)			Stop codons (%)		
	AUG	GUG	UUG	UAA	UAG	UGA
AUG	82	13	5			
GUG	82	14	4			
UUG	81	14	5			
UAA				56	19	25
UAG				54	18	28
UGA				49	18	33

Bs, *Bacillus subtilis*; Oth, other.

Table 3. Average energy of the mRNA secondary structure near mutated start and stop codons (kcal/mol)

	Start codons			Stop codons		
	AUG	GUG	UUG	UAA	UAG	UGA
AUG	–4.54	–5.11	–4.93			
GUG	–4.88	–5.50	–5.39			
UUG	–4.71	–5.50	–5.30			
UAA				–5.40	–5.45	–5.70
UAG				–5.87	–6.00	–6.14
UGA				–5.89	–5.89	–6.40

True codons are in columns and mutated codons are in rows. For example, the cell in column UUG and row AUG contains the mean energy of the 50 bp sequence surrounding the UUG stop codons, when they are mutated *in silico* to AUG. Values on the diagonal correspond to the mean energy surrounding the original codons.

A-richness around the start constrains codon usage bias

Nucleotide distribution before the start codon reveals peaks of high frequency of G and A at the corresponding RBS positions, and low frequency of C (Fig. 1). Nucleotide distribution approaches that of the average coding region after position 30 in the genes, and A is particularly over-represented at the beginning of the genes. We have built contingency tables for each amino acid, comparing the average codon usage in the gene with the codon usage from positions +2 up to +10. This analysis revealed significantly different codon usage for nearly all amino acids, always in the sense of increasing the number of A terminating codons (and U in doublet codons) (Table 4). The analysis of the export/non-export and *B.subtilis/E.coli* sets revealed similar results, with over-representation of A terminating codons (and U to a much lesser extent), independently of the nature of the amino acid (data not shown).

Avoidance of mRNA structure may cause nucleotide and codon usage bias at the start

RNA structure is unstable near the start (Fig. 1), and the stability is significantly lower at the start codon than at the RBS. The energy near the start is significantly different between class 1 genes and classes 2 and 3, with an absolute difference of 10% between the averages (–4.8 and –4.4 kcal/mol respectively). De Smit and van Duin (3) have shown that the translation efficiency depends on the equilibrium between the strength of the RBS–ribosome interaction and the mRNA structure. Since mRNA structure is minimal at the start codon, not at the RBS, one may think that the equilibrium between the codon–anticodon interaction and the mRNA structure should be added to this model. This is consistent with the experimental finding that in *E.coli* and *B.subtilis* stronger RBS severely diminish the differences of degree of expression between different start codons for a given gene (6), and to the finding that leaderless transcripts without secondary structure can attach correctly to the ribosome (43).

If unstable mRNA structure at the RBS (and probably at the start codon) is essential for translation initiation, one may suppose that nucleotide and codon usage bias at the start is also a function of this constraint, particularly since both biases act towards over-representation of A (or U in 2-codon amino acids), which is the most efficient way of reducing mRNA secondary structure. This trend is also independent of the choice of the start codon, of the RBS strength, of the nature of the encoded protein and of its degree of expression. It is also mostly independent of over-represented amino acids since some of these are A-poor. There is no apparent relation between such

Table 4. Codon usage bias by contrast to the typical codon usage near the start and stop codons

		After start	Before stop
A	+	A, G, P, T, V, R, L, S, Q, E, K, I	A, G, P, V, R, L, S, Q, I
	−		E
C	+		P
	−	A, G, T, V, R, L, S, F, I	A, G, V, R, L, N, H, F, Y, I
G	+		G, V, R, E
	−	A, P, T, V, R, L, Q, E, K	L, Q, P
U	+	G, T, S, F	L, N, H, F, Y
	−	I	G, R, S, I

The significance of the contrast between two sets (e.g. alanine codons after the start to average alanine codons usage), was assigned through the use of a contingency table. The outliers associated with each variable were taken as significant when their contribution to the χ^2 value was >3.8. For example, for alanine (codons GCN) there is a preference (+) for A-richness (in comparison to the typical codon usage bias in the inner gene) and C-avoidance (–) near the start and stop codons.

unidirectional bias towards A and selection for lower translation levels (44), as experimental results have shown (8).

The spacer between the RBS and the start codon is A-rich and C-poor, whereas in *E.coli* is A-rich and G-poor (21). This also reinforces the idea that these biases act to reduce mRNA secondary structure since the absence of S1 in *B.subtilis* probably implies requirements of less stable structure, which following a G-rich region (due to the RBS), is achieved by avoiding C. In *E.coli* where these requirements are less stringent the typical C-avoidance of transcribed sequences is the dominant factor.

Finally, mRNA structure stability decreases up to position +30 after the start where it becomes close to that of the middle of genes (Fig. 1), confirming previous results (23). This A-rich unstructured region marks the known limit of the interaction of the ribosome and the mRNA at the initiation phase (45).

Oligonucleotide bias in the spacer is typical of coding regions

The spacer (region between the RBS and the start codon) and the average intergenic regions present large differences in terms of word usage even after removing the RBS. The correlation between dinucleotide bias in the spacer and in the genes is 0.98, but between the spacer and the intergenic regions it is only 0.60. For trinucleotide bias the first correlation is 0.88 and the second is 0.20. For tetranucleotides the latter is not significantly different from 0 (data not shown). This observation reinforces our previous analysis that part of these biases are caused by the interaction with RNA polymerase in transcription and with the ribosome complex in translation (37).

Amino acid bias at the start is mostly similar among different functional classes

Nine amino acids present overall biases at the beginning of proteins (without predicted signal peptides) by comparison with the average protein composition. All four over-represented amino acids are hydrophilic (Lys, Asn, Gln, Ser), whereas the five under-represented are all hydrophobic (Ala, Gly, Leu, Pro, Val). A comparative analysis of the export and non-export sets (Materials and Methods), revealed that biases are not due to possibly unrecognised signal peptides. In fact, as shown in Figure 2, serine and glutamine turn out to be much more biased in the non-export set, all under-represented amino acids are similarly biased in the two sets, with the exception of alanine. Threonine and tyrosine are over-represented in the non-export set. Most biases are stronger at position +2 but extend up to position +10 and are position dependent. These amino acid biases are roughly similar in *B.subtilis* and *E.coli* proteins (Table 5).

This analysis reveals that biases usually attributed to peptide signals [e.g. lysine-rich sequences (46)] are in fact general. Moreover, some biases (e.g. over-representation of serine) are stronger for the central metabolism proteins. It is not possible at this stage to indicate if we are in the presence of a single protein bias, to which a particular signal peptide bias is over imposed, or if there are different biases (but following some general trends as can be inferred from Table 5) for different protein types. Clearly these biases cannot be fully explained by the N-end rule for eubacteria (47), since they extend well beyond the first amino acid. Recently, an analysis of the proteome of all completely sequenced genomes has shown that some of these biases are general for most bacterial species (48).

mRNA structure is more stable than expected in the middle of genes

The average energy of the secondary structure of inner gene sequences randomised in nucleotides is –6.1 kcal/mol, which decreases to –6.6 kcal/mol if randomisation is made keeping trinucleotides composition. Both of these values are significantly different from the observed average of –6.9 kcal/mol (using the Tukey–Kramer test) (Fig. 1). Genes of FCA classes 1 and 2 have equivalent levels of mRNA structure, but class 3 genes show significantly less stable secondary structures, which is probably caused by the high A+T content of these genes (64% in class 3 genes for 57% in the genome).

Table 5. Amino acid bias at positions 2 to 10 (position 1 is Met) of four different protein sets: all *B.subtilis*, all *E.coli*, *B.subtilis* class non-export and *B.subtilis* class export

	χ^2	*B.subtilis*		*E.coli*		*B.subtilis* non–export		*B.subtilis* export	
		+	–	+	–	+	–	+	–
2	**701**	**K,N,S**	F,I,V,C,G,H,L	**K,M,N,S,T**	D,V,Y,C,E,G,H,L,W	**K,N,T,S**	**V,G,C,W,L,H**	**K,N,R,S**	F,I,M,V,Y,G
3	**493**	**K,N,Q,R,T**	F,A,P,G,L	**I,K,M,N,Q,T**	A,P,R,V,G,L	**K,N,Q**	P,V,G	**K,N,R,T**	A,P,V,G
4	**238**	**K,N,Q,R**	**A,G**	**I,K,N,S**	A,D,R,V,E,G,H	**K,Q**		**K,R**	A,P,G
5	**166**	**K,Q**	**A,G**	**I,K,N,T**	**A,E,G**	**K**	**G**	**K,R**	**A,G**
6	**99**	**K**	**A,E,G**	F,I,K,T	A,D,R,V,E,G	**I,W**	**G**	**K,N,R**	G
7	**83**	**K**	**A,G**	**I,K**	**E,G**	**I,Q,R**	**G**	K,R	P,G
8	**50**		**G**	**I,K**	**A,R,E,G**	**D**	**F**		**A**
9	43		A	I,K	Y,E			**K,R,H**	
10	38				**E**		**Y**		

+, over-represented amino acids; –, under-represented amino acids.
Column 2 indicates the χ^2 value between the distribution of amino acids at all positions and the distribution at each position, in the *B.subtilis* set. After position 8, the bias becomes not significant at the 1‰ level. In all columns bold characters indicate that the bias is significant at the 1‰ level (1% otherwise).

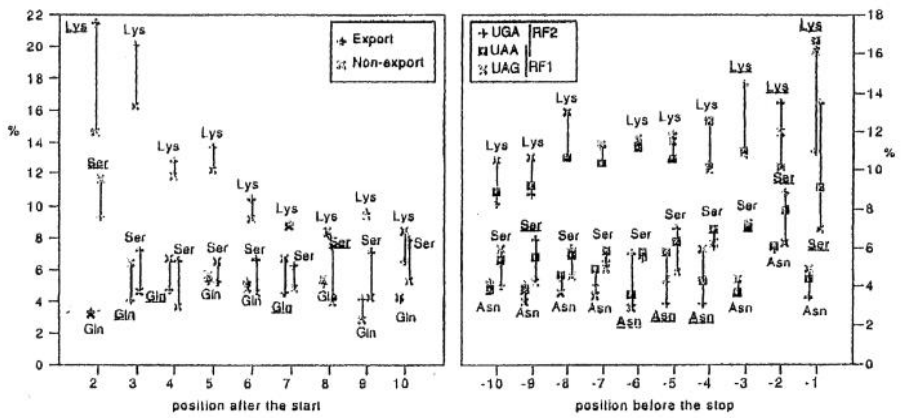

Figure 2. Relative abundance of lysine, serine and glutamine at N-terminus of proteins of the two sets, export and non-export (left). Relative abundance of lysine, serine and asparagine at the C-terminus as a function of the stop codon (right). Values in bold and underlined indicate significantly different means between the sets.

UAA usage is dependent on G+C content and on gene expression level

The stop codon usage changes more considerably than the start among prokaryotes (Table 1). There is also little conservation of the stop codon between homologues (Table 2). The use of the codon itself is significantly correlated with the G+C content of the genome, negatively for UAA (Spearman's rank of –0.70) and positively for UGA (0.76). For UAG the correlation is not significantly different from zero. The order of relative frequencies of stop codons in *B.subtilis* is UAA > UGA > UAG, and class 2 genes use almost exclusively UAA (Fig. 1). Preliminary analysis revealed similar results for *E.coli*, but not *B.subtilis*, probably for lack of enough data at the time (52 genes) (49).

mRNA structure stability requirements bias codon, nucleotide and stop codon usage

Excluding the structure of the rho-independent terminators, we find that mRNA structure is less stable near the stop codon (Fig. 1). Mutating UGA either to UAA or to UAG leads to more stable structures, and the reverse happens when mutating these to UAA (Table 3). The analysis of the structure covering the stop codon reveals a significant energy decrease in the sense UAA > UAG > UGA.

There is a clear over-representation of A near the stop codon, though less important than at the start. Correspondingly codon usage is also biased towards an increase in A independently of the stop codon, with few exceptions (Table 4).

Reduction of mRNA structure at the stop codon may partly explain all these biases. Since the major event in the termination is the interaction between the stop codon and the release factor (50), it is likely that readthrough is better avoided by lowering mRNA structure around the stop codon. This may explain the preference for UAA, particularly in highly expressed genes: not only UAA is read by both RF [with the same efficiency as the other stop codons (51)], but also avoids mRNA structure. Additionally, the interaction stop codon/RF is facilitated by an A-rich surrounding that decreases mRNA structure stability.

Nucleotide bias after the stop are RF and stop codon dependent

The six positions following the stop are very A+U rich (Fig. 1). Though A and G at positions +1 to +3 seem to favour readthrough (41), and have been found to be under-represented in a small sample of *B.subtilis* genes (24), we found 40% of A at the +1 position. One may speculate that A at this position is only significantly disturbing in highly expressed genes, where U becomes dominant (56% in class 2 against 27% in all genes), possibly because it minimises readthrough [as observed for RF2 in *E.coli* (52,53)].

The UGA stop is mainly followed by A or U, whereas UAG is followed by A or G. This fourth base bias may be partially explained by the fact that GT and AG are highly avoided dinucleotides in *B.subtilis*, whereas GG is much less so, and AT and GA are over-represented (37). The patterns observed for UAA are very close to the ones found for UGA, in what concerns G and U at position +1. In the *E.coli* RF2 system UAGG is more efficient than UAGU (and UAAU and UGAU more efficient than UAAG and UGAG) (52). This probably means that the larger signal hypothesised for the stop codon (18) is RF specific. However, an explanation for these biases based solely on the RF specificity is not fully satisfactory since UAA exhibits the most extreme positive bias of A everywhere and the most negative bias of C at position +1 (Fig. 3) whereas one would expect intermediate behaviour for this stop codon since it is recognised by both RF.

Dinucleotide bias after UGA is highly correlated with dinucleotide bias in non-coding regions (correlation of 0.91), but this is not the case for UAG (0.14), which may indicate that bias at this level may be discriminative. Biases of larger oligonucleotides are similar for all stop codons and reflect the average distribution of intergenic regions.

Amino acid bias before the stop is strongly RF specific

The few experimental results available for *B.subtilis* indicate trends of amino-acid-dependent readthrough rates similar to *E.coli*, with lysine providing good stops and proline and threonine bad ones (54). Our results point out that a larger number of amino acids may be particularly advantageous (e.g. arginine, asparagine and serine; Table 6), and that these trends are partially release-factor specific.

Amino acid bias at the stop is mostly restricted to the last two positions although arginine and lysine are over-represented within the last 10–15 amino acids (Table 6). Serine is much more over-represented in UGA-ending genes, whereas lysine is over-represented in the last position only if the stop is UAG, indicating an RF-specific bias (Fig. 2). Intriguingly, glycine at position –4 is more than three times more frequent in UGA

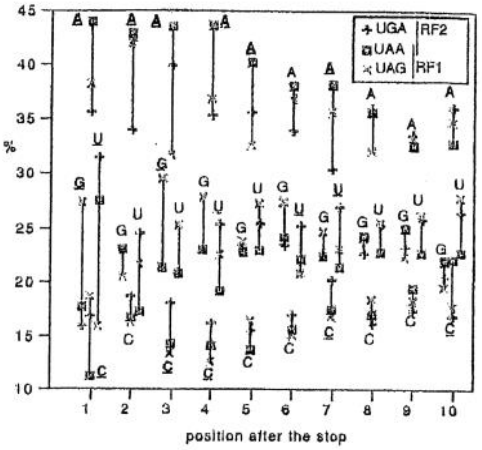

Figure 3. Relative abundance of nucleotides after the stop codon as a function of stop codons. Values in bold and underlined indicate significantly different means between the sets.

terminating genes than in UAG. Some of the remaining amino acids also show position-dependent differences in terms of stop codon preferences (Table 6). UAA genes are biased as UAG for lysine and isoleucine and as UGA for serine. Therefore, the preferences of UAA in terms of amino acids cumulate the two main biases.

If amino acid bias is due to interaction between the nascent peptide and the ribosome (19), then this interaction should include (and discriminate among) the release factors. In fact, differences of up to 30 times in stop readthrough were found by changing the penultimate amino acid, though the difference was much less important when the stop is UAG rather than UGA (55).

Amino acid biases at both termini of proteins are similar

Our results point to the existence of similar general amino acid biases at both termini of proteins. The globally biased amino acids at both extremities are the same, namely over-representation of hydrophilic (Lys, Arg, Ser and Asn) and under-representation of hydrophobic amino acids (Ala, Ile, Gly, Leu). These biases are strongest at the extremities of proteins, extend along 10 bases, and are partially position dependent (Fig. 2). An A at the second position of the codon codes for hydrophilic residues, and therefore A-richness to avoid stable mRNA structures may cause an over-representation of hydrophilic residues. However, it does not fully explain the results since codons for biased amino acids such as arginine and serine do not have an A at position 2 and are not particularly A-rich.

CONCLUSION

Since translation is the most energetically demanding process of the bacterial cell, one should expect it to be highly optimised. Optimisation is a function of the element's importance and the environmental context (2), and is particularly relevant at high growth rates. This optimisation proceeds at four levels: initiation,

Table 6. Amino acid bias at the C-terminal positions −10 to −1 of four different protein sets: all *B.subtilis*, all *E.coli*, *B.subtilis* UAG-ending genes and *B.subtilis* UGA-ending genes

	χ^2	*B.subtilis*		*E.coli*		*B.subtilis* UAG		*B.subtilis* UGA	
		+	−	+	−	+	−	+	−
−1	**356**	**K,R,S**	**I,P,T,V,Y,G,L**	**K,Q,R,E,H**	**I,M,P,T,V,L**	**K,R,E**	**I,P,G**	**K,R,S**	**P,T,L**
−2	**153**	**K,N,Q**	**I**	**K,R**	**F,I,P**	**K,N,R**	**P**	**K,N,R,S**	**P,L**
−3	**107**	**K,E**	**D,G**	**K,R,E**	**F,A,G**	**K**	**V,G**	**K**	**P**
−4	**137**	**K,R,E**	**I**	**K,R,E**	**I,T,L**	**K,R**	**G**	**K,R**	**I,E,G,L**
−5	**99**	**K,N,R,E,W**	**A**	**K,R**	**G**	**K**		**K,R**	**A,I,E,G**
−6	**77**	**K,R**	**A,G**	**K,R**		**K**		**K,N,R**	**A,E,L**
−7	**80**	**K,R,E**	**V**	**K,R**	**S,G**	**K**		**K,R**	**E**
−8	**79**	**K**		**K,R**	**G**	**K**		**K,R**	
−9	**52**	**K**	**G**	**K,R**	**G**	**K**	**P**	**R**	**F**
−10	**66**	**K,R**	**G**	**K,R**	**F,N**	**D,K**		**Q,R**	**F**

+, over-represented amino acids; −, under-represented amino acids.
Column 2 indicates the χ^2 value between the distribution of amino acids at all positions and the distribution at each position, in the *B.subtilis* set. In all columns bold characters indicate that the bias is significant at the 1‰ level (1% otherwise).

codon usage bias in elongation, termination and mRNA stability. In this work emphasis was given to the initiation and termination, since many studies have been published on the subject of codon usage (29,30), and the study of mRNA degradation requires the analysis of polycistronic units, unavailable for *B.subtilis* at this time (25,27).

Our analysis was first divided into signals, compositional biases and structures. Our results indicate that compositional biases are most determinant at the amino acid level, since the remaining can be explained by extension of signals (e.g. extended stop codon) or mRNA structure avoidance. Since we demonstrate that the 'downstream box' is not statistically significant and most other proposed signals were found to be rather system-specific or doubtful (4), we are inclined to believe that there are only three general signals at the translation level: start and stop codon and the RBS. The remaining patterns are a consequence of trends acting to diminish mRNA structure or to discriminate between different variants of a signal.

The competition between the ribosome–RBS interaction and the mRNA structure results in nearly 'all or none' expression, leaving almost no room for regulation at the purely translation level (i.e. excluding regulation at the full transcript level) (3). Therefore, all moderately or highly expressed genes should have a good combination of low mRNA structure, good RBS and efficient start codon. On the other hand, since the system has a certain degree of freedom to mutate (between the RBS, the start codon and the mRNA structure), a positive mutation (e.g. reducing the structure or strengthening the RBS), may compensate a negative one (e.g. a mutation on an AUG). This high degree of flexibility of the system may explain the lack of conservation of the start codons among homologous genes.

It is intriguing to find that UAA and UGA abundances are highly correlated with G+C content whereas UAG is not. The change of the nucleotide after UGA results in readthrough increase of up to 30 times, whereas for UAG this difference is three times smaller (52). The same selective dependency was seen for the amino acids preceding the stop codon (55). If this means that RF2 termination is more dependent on the biases surrounding the stop codon, then UAA should follow UGA patterns more closely. This in turn should imply that UAA→UGA transitions should be less deleterious than UAA→UAG. However, not only is this not observed in the analysis of the homologues (Table 2), but the amino acid bias (where UAA follows more closely UAG) follows the exact opposite trend. For the moment this remains an open question.

ACKNOWLEDGEMENTS

We thank T. Nogueira for comments on the manuscript. This work was partially funded by the EU grant Biotech BIO4-CT96-0655. E.R. acknowledges the support of PRAXIS XXI, through the grant BD/9394/96.

REFERENCES

1. Jacques,N. and Dreyfus,M. (1990) *Mol. Microbiol.*, **4**, 1063–1067.
2. Andersson,S.G.E. and Kurland,C.G. (1990) *Microbiol. Rev.*, **54**, 198–210.
3. de Smit,M.H. and van Duin,J. (1994) *J. Mol. Biol.*, **235**, 173–184.
4. Draper,D.E., (1996) In Curtiss,R., Ingraham,J.L., Lin,E.C.C., Brooks Low,K., Magasanik,B., Reznikoff,W.S., Riley,M., Schaechter,M. and Umbarger,H.E. (eds), *Escherichia coli and Salmonella: Cellular and Molecular Biology*. ASM Press, Washington, DC, pp. 902–908.
5. Vellanoweth,R.L., (1993) In Sonenshein,A.L., Hoch,J.A. and Losick,R. (eds), *Bacillus subtilis and other Gram-positive Bacteria*. American Society for Microbiology, Washington, DC, pp. 699–711.
6. Vellanoweth,R.L. and Rabinowitz,J.C. (1992) *Mol. Microbiol.*, **6**, 1105–1114.
7. Sprengart,M.L. and Porter,A.G. (1997) *Mol. Microbiol.*, **24**, 19–28.
8. Dreyfus,M. (1988) *J. Mol. Biol.*, **204**, 79–94.
9. Bulmer,M. (1988) *J. Theor. Biol.*, **133**, 67–71.
10. Sprengart,M.L., Fatscher,H.P. and Fuchs,E. (1990) *Nucleic Acids Res.*, **18**, 1719–1723.
11. Plat,T. (1998) In Grunberg-Manago,M. (ed.), *RNA Structure and Function*. Cold Spring Harbour Laboratory Press, Cold Spring Harbor, New York, NY, pp. 541–574.