RELEVANCE OF LYSINE ABUNDANCE AT SECOND CODON POSITION IN *ESCHERICHIA COLI*

Talal Ali

300309840

Supervisor: Dr. Xuhua Xia

Research report submitted to the Biomedical Sciences Program in partial fulfillment for the requirements for the course BIM4009

University of Ottawa

Ottawa, Ontario, Canada

May 2026

© May 2026, Talal Ali

ABSTRACT

Understanding nucleotide distribution patterns is key to improving mRNA translation efficiency, which can serve many functions, such as understanding cell function and mRNA vaccine optimization. In the *Escherichia coli* genome, a preference for lysine in the second codon position (P1') is observed. One hypothesis state that lysine is abundant due it is protecting the N-terminal methionine from N-terminal methionine excision. The second hypothesis is that lysine is only occurring incidentally because a poly(A) stretch is favoured in the sequence surrounding the start codon due to it easing ribosome binding by reducing messenger RNA secondary structures. To determine the function of this lysine, the codon frequencies of lysine and alanine were calculated, as well as the nucleotide frequencies of *E. coli* genes and other bacteria. AAA was more common than AAG, however, GCA and GCU was preferred over other codons for alanine. *E. coli* was found to have an A-rich sequence downstream of the start codon, especially when compared to slow reproducing bacteria. These results suggest that mRNA secondary structure is responsible for lysine abundance in the second codon position for *E. coli*.

ACKNOWLEDGEMENTS

I thank my supervisor Dr. Xuhua Xia for helping me conceive the idea and methodology for the thesis as well as his valuable feedback and expertise. I would also like to thank my cohort Ann-Sophie Gupta, Christina Bortnowschi, Safaa Hadroune, Sarah Benomar and Omer Asghar for their valuable feedback and time.

STATEMENT OF AI USE

I did not use AI to generate any text, figure, or table in the thesis.

DATA AVAILABILITY

All data was acquired from the National Center for Biotechnology Information (NCBI) from the accession codes mentioned later in Table 1.

TABLE OF CONTENTS

TITLE PAGE I

ABSTRACT II

ACKNOWLEDGEMENTS III

STATEMENT OF AI USE III

DATA AVAILABILITY III

TABLE OF CONTENTS IV

INTRODUCTION 1

A) The methionine preservation hypothesis 5
B) The ribosome binding hypothesis 8
C) Other ideas 9
D) Objective 10

METHODS 11

RESULTS 13

A) Codon distribution of lysine at P1' 13
B) Codon distribution of alanine at P1' 14
C) Adenine distribution of *E. coli* 15
D) Different bacterial species' position weight matrix scores for adenine 18
E) Intragenome *E. coli* position weight matrix scores comparison for adenine 20

DISCUSSION 22

A) Codon distribution of lysine at P1' 22
B) Codon distribution of alanine at P1' 22
C) Adenine distribution of *E. coli* 23
D) Different bacterial species' position weight matrix scores for adenine 24
E) Intragenome *E. coli* position weight matrix scores comparison for adenine 25

CONCLUSION 25

REFERENCES 27

APPENDIX 33

INTRODUCTION

A genome's major patterns can reveal information about strategies the organism uses on a molecular level. Deoxyribonucleic acid (DNA) is comprised in part by adenine (A), cytosine (C), guanine (G), and thymine (T) nitrogenous bases, and in ribonucleic acid (RNA), thymine is replaced by uracil (U). In general, during transcription, segments of DNA called promoters recruit RNA polymerase, which transcribes the DNA into messenger RNA (mRNA). This mRNA contains the start codon, a set of 3 nucleotides AUG, and a stop codon, which can take the form of multiple codons and is specific depending on species such as *Escherichia coli* (See *Supplementary Figure 1*). The mRNA is then taken up by a ribosome, which, with the help of transfer RNA (tRNA), synthesizes a polypeptide chain. Each set of three nucleotides in the mRNA corresponds to an amino acid, with the specific correspondent being dependent on the species. This polypeptide is a polymer of amino acids joined by peptide bonds, with one end being the α-amino group of the amino acid backbone, called the N-terminus, and the other end is the α-carboxyl group called the C-terminus.

In bacteria like *E. coli*, as well as eukaryotic symbionts, the start codon AUG adds a form of methionine called N-formylmethionine at the N-terminus, making it the most frequent N-terminal amino acid [1] before processing. This modified amino acid is made by adding formyl to the amino acid attached to its tRNA, methionyl-tRNA, by a methionyl-tRNA formyltransferase (FMT), this enzyme has long been isolated, initially referred to as N^{10}-formyltetrahydrofolate:methionyl soluble ribonucleic acid transformylase thanks to its use of 10-formyltetrahydrofolate as a formyl donor [1,2]. For prokaryotes like *E. coli*, this amino acid is

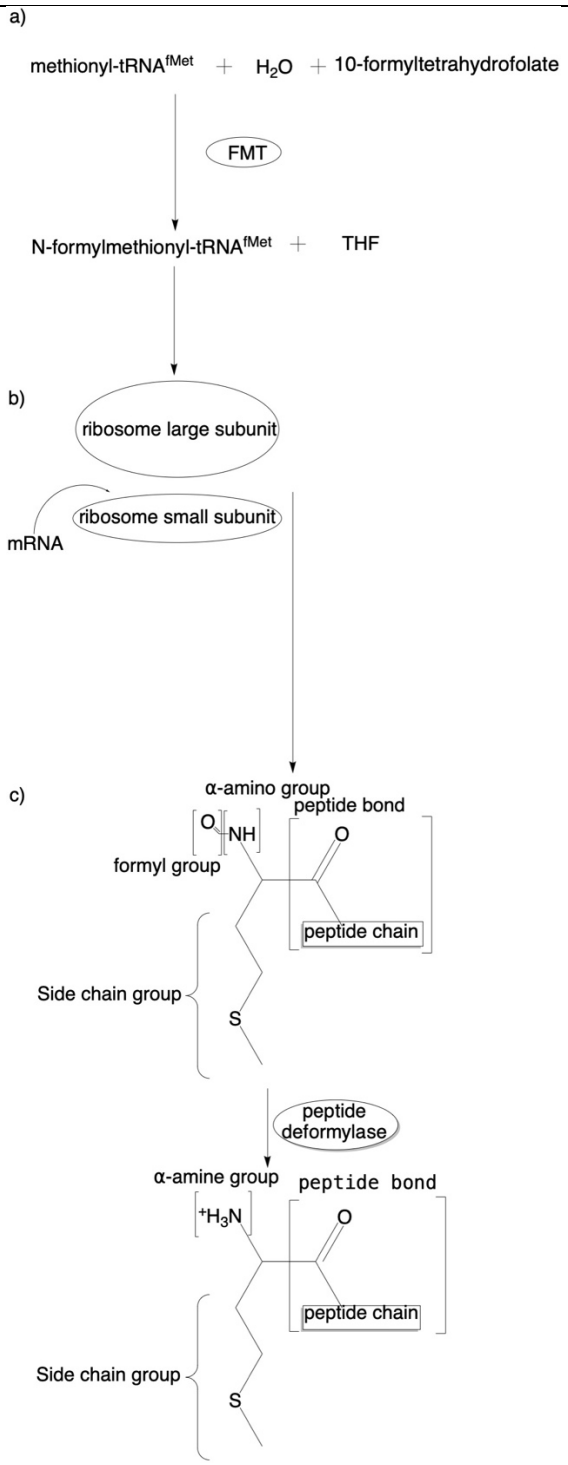


Figure 1. Schematic of co-translational processing of methionine before excision

(a) Methionyl-tRNA is turned into N-formylmethionyl-tRNA and tetrahydrofolate
(b) Ribosome adds N-formylmethionine as N-terminal amino acid (translation)
(c) Formylmethionine is changed by peptide deformylase, leaving N-terminal methionine, formyl group is removed, allowing further processing

often processed by a peptide deformylase, leaving N-terminal methionine for further processing [3] if needed, since not all methionine amino acids are excised in *E. coli* due to some genes needing one at the N-terminal position. Figure 1 presents this process. Once the formyl group is no longer attached to the α-amino group, the methionine may undergo N-terminal methionine excision (NME). In the context of methionine excision, the amino acid chain P1-P1'-P2', etc. can be designated as the substrate. The site of cleavage would be between P1 and P1', where P1 is the methionine to be excised. This terminology was initially introduced by Schecter & Berger in 1967 [4,5]. The amino acid at the second position after the start codon is therefore referred to as P1', and its corresponding amino acid in the active site of the enzyme can be referred to as S1'.

A transcription unit is a sequence in DNA that codes for a single RNA, normally containing a promoter, RNA code and a terminator. It is the more general DNA segment representing a polycistronic operon that would have many genes transcribed onto a single mRNA. For example, on a single strand of mRNA, there could be multiple co-regulated genes with a single promoter. The ribosomes would take up the mRNA containing multiple genes and there would be a better stoichiometric balance in the number of each protein translated. Assuming ribosomes need a predictable number of subunits for every gene translated, it would make sense for them to be clustered in a transcription unit.

Among gram-negative bacteria, *E. coli* genomes are some of the most studied in genetics, their structure and patterns can provide insights into many genetic mechanisms used by organisms. The circular genome of a particular strain, *Escherichia coli* str. K-12 substr. MG1655, is about 4.6 million base pairs long, and has many studies done for annotation [6,7]. MG1655 was first isolated in 1922 and was eventually chosen as a representative for the K-12 wild-type

due to having fewer point mutations and chromosome segment inversions compared to alternative strains [8,9]. Despite these annotations, the function of nucleotide patterns in the genome is not yet fully understood. In *E. coli*, the most common amino acid at the P1' position is lysine at 16.7%, and it has been suggested that this preference for lysine changes to serine in lower eukaryotes to alanine in higher plants and animals [10]. There is a strong likelihood that lysine is preferred at this position, since among 20 basic amino acids, a random distribution would be closer to 5%. Another study by Bivona et al. in 2010 suggested that the codon in the P1' position greatly affected protein expression [11]. The same investigation showed that the preference at this position was for amino acids rather than codon usage due to many amino acids in the second position for highly expressed genes showing no preference for specific codons among synonymous codons; however, a clear preference among codons for lysine at this position was observed.

E. coli can double every 20 minutes under ideal aerobic and nutrient-rich conditions in the laboratory [12], making it a fast-growing bacterium. Such rates of reproduction would presumably require an equally efficient translation process. A study by Xia in 2026 demonstrated that slow-growing and fast-growing bacteria differ in translational machinery in the number of tRNA and ribosome RNA operons [13]. If there are patterns of nucleotide distribution that can be attributed to translation efficiency, such a pattern should be detectable on the *E. coli* genome. Similarly, bacteria that reproduce more slowly should not have this nucleotide distribution pattern. Translation efficiency is divided into initiation, elongation, and termination efficiency, with a progressively decreasing contribution to overall translation efficiency. There are many possible mechanisms of modulating translation efficiencies, among them are mRNA secondary structure contributions and affinity to the Shine-Dalgarno sequence [14]. More complex

mechanisms, such as certain amino acid sequences interacting with ribosomes to stall their own elongation, frameshifting for different termination efficiency, tRNA and ribosome availability and many more [13,15,16]. All these factors can contribute to the patterns observed, the question is what causes the abundance of lysine at the P1' position?

A) The methionine preservation hypothesis

There are multiple competing hypotheses explaining lysine abundance after the N-terminal methionine. One of the leading hypotheses is regarding NME in *E. coli.* As previously discussed, N-formylmethionyl-$tRNA^{fMet}$ is formed by the activity of FMT on methionyl-$tRNA^{fMet}$. After translation, by the activity of peptide deformylase, the formyl is removed and methionine residue remains in place as presented in Figure 1. For any given proteome, up to 80% of proteins undergo NME [17,18], including prokaryotes. Generally, NME can use two types of methionine aminopeptidase (MAP), however, for most prokaryotic organisms such as *E. coli*, only one form of MAP is used after peptide deformylase has removed formyl from the N-terminal amino acid [17,19]. Since NME is such a prevalent occurrence in organisms, genome-wide trends are easier to observe. In addition, the fact that *E. coli* only has one form of MAP, meaning it a simpler model to study compared to eukaryotic organisms.

For the MAP enzyme, the portion corresponding to the P1' of the protein called the S1' binding pocket seems to work better with the smaller and less bulky amino acid side chains of alanine, cysteine, glycine, proline, etc. at the P1' position, while the ones with larger side chains like lysine and arginine seemed to prevent NME activity [17,19–21]. In addition, charged amino acids are more likely to hinder NME [22], making small and non-polar amino acids at the P1' position a good indicator of methionine excision. Therefore, having a lysine at the P1' position

would exclude the protein from NME, this N-terminal methionine may be important for the function of many proteins. One example of proteins that would not require a smaller amino acid to cleave N-terminal methionine are secretory proteins. They generally do not undergo NME, instead their precursor signal peptide is removed by leader peptidase I during the secretion process [23]. In fact, in a study by Zalucki et al. in 2007, the link between NME and lysine abundance at the P1' position was explored, due to the change in charge and the change in amino acid on the P1' position not affecting protein export, and the fact that the AAA codon is preferred among lysine codons, they concluded that high translation initiation efficiency caused the abundance of lysine at the P1' position. For the purpose of this report, we can generalize their N-terminal methionine preservation hypothesis.

The hypothesis is as follows: lysine abundance at the P1' position is caused by the need to preserve some N-terminal methionine in proteins, since having lysine at that position prevents NME from occurring. This hypothesis is more general than the previously mentioned study, as it doesn't inquire the reason why methionine needs to be shielded from MAP activity.

The precise catalytic mechanism of NME is also worth looking at, especially since the actual mechanism is newly being discussed. During NME, N-terminal methionine is still attached to the peptide chain, the polypeptide acts as a substrate for MAP, since it is a metalloprotease, metal ions (Co^{2+}, Mn^{2+}, Fe^{2+}) and water would turn into a reactive hydroxide, allowing for peptide bond cleavage [24–27]. A simplified version of the mechanism is shown in Figure 2. Essentially, the length and the charge of the amino acid at P1' control the activity of MAP.

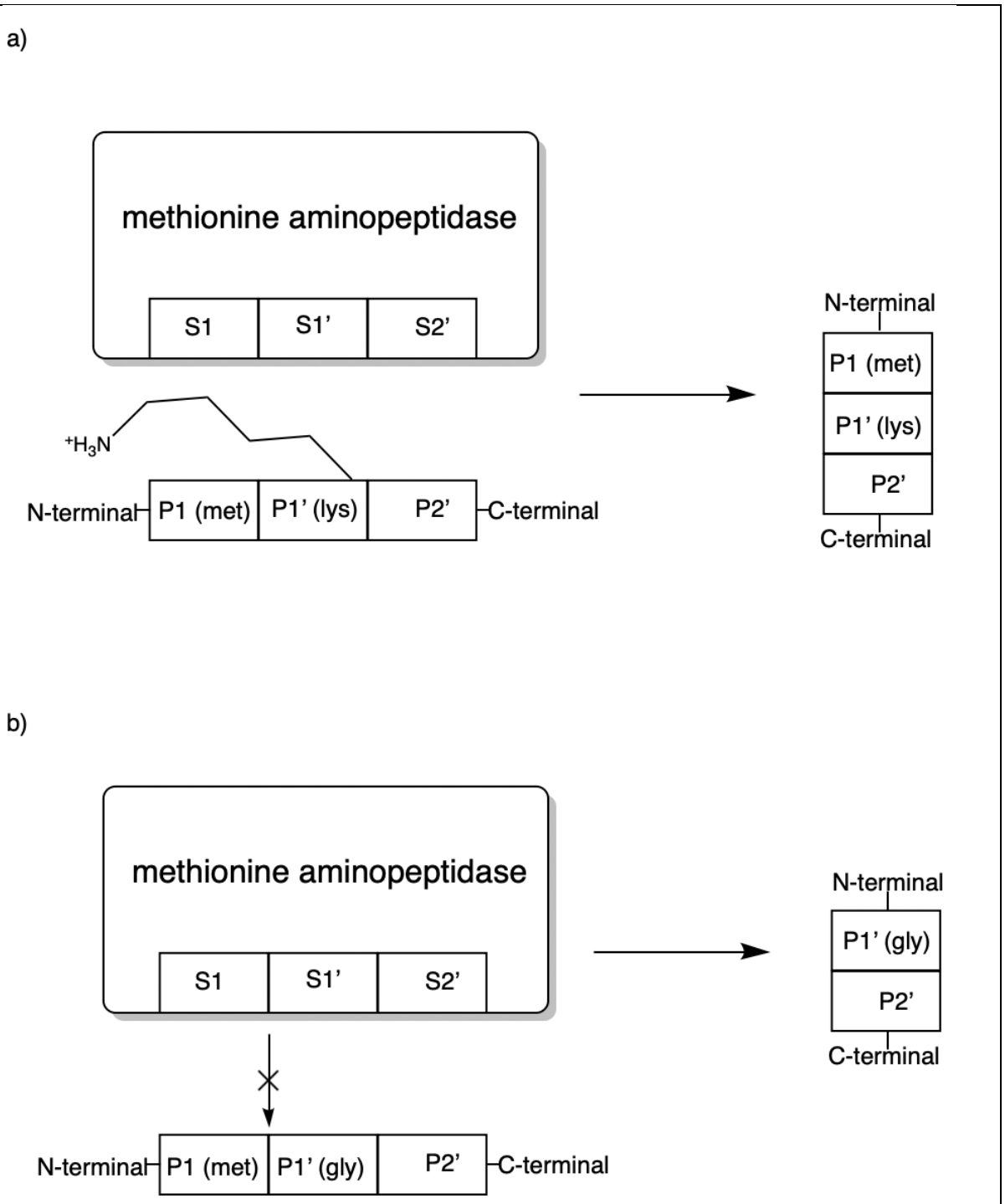


Figure 2. Representation of methionine aminopeptidase activity with different P1' amino acids

(a) Longer amino acids like lysine prevent excision of methionine, S1-S2' represent the active site
(b) Shorter amino acids like glycine allow excision of methionine, S1-S2' represent the active site

B) The ribosome binding hypothesis

On the other hand, the distribution of nucleotides within protein-coding sequences may be related to the secondary structure of mRNA [28]. When it comes to the 5'UTR region of mRNA, complex secondary structures can impede the function of ribosome binding to the cap, reducing translation efficiency [29]. Due to the guanine and cytosine pairs having 3 hydrogen bonds, they can form more stable and complex secondary structures compared to the adenine and uracil pair, which only have 2 hydrogen bonds [30]. This means that GC-rich protein coding sequences would form more secondary structures compared to AU-rich sequences.

Considering the Shine-Dalgarno sequence is required for more efficient ribosomal binding, the 5' UTR region of mRNA cannot be mutated excessively. This sequence is highly conserved, especially in highly expressed genes. However, many of the nucleotides in the sequence surrounding it can mutate without disrupting the Shine-Dalgarno sequence. Only looking at the sequence immediately downstream of the start codon, there is no such sequence or promoter, meaning the likelihood that mutations don't disrupt gene function is higher than those highly conserved segments. Although more evidence later accumulated, even in early studies, such as one by Hall et al. in 1982, the role of the secondary structure of mRNA in translation initiation was being discussed [31]. It wasn't until Zalucki et al. did their paper in 2007 that lysine abundance at the P1' position was specifically linked to mRNA secondary structure [23]. Nevertheless, this initial study did not explore this dimension of the topic in as much detail. After seeing that AAA was responsible for a strong bias in lysine residues at the P1' position, they postulated that there was a non-coding aspect to this bias. There are multiple forms secondary structures can take in mRNA, for example, G-quadruplex, hairpin, etc. [32]. However, it has been observed that there is an increase in secondary structure formation for G and C-rich mRNA

[33], and that increased secondary structure at the 5' UTR is inversely proportional to translation efficiency [34]. This leads perfectly into the ribosome binding hypothesis.

The hypothesis explicitly stated for the purpose of this report is as follows: lysine amino acid is not the relevant factor in its abundance; it only occurs incidentally as a consequence of being represented by AAA and AAG codons. The relevant part is the frequency of adenines near the start codon, as more adenines would reduce the secondary structure of mRNA carrying these codons near the start of the gene, which would allow the ribosome to attach with the mRNA more efficiently. A poly(A) stretch would reduce the secondary structure, and a poly(A) necessarily implies having AAA codons, which would make it more likely to have lysine at the P1' position.

C) Other ideas

Lysine contains an amino group ($R\text{-}NH_3^+$) that is often positively charged, giving the protein the characteristic of a variable and controllable charge. This can be done by the acetylation of lysine, through the transfer of an acetyl group by the acetyl coenzyme A to the amine group on the lysine side chain [35]. This leads to the lysine losing its positive charge, leading to an overall decrease in the charge of a given protein. In the case of histones, for example, positively charged lysine amino acids help bind negatively charged DNA, leading to decreased accessibility to the genome, and the acetylation of lysine can reverse this, unwrapping the DNA [36–38]. However, this idea doesn't hold up to scrutiny, since such a strategy would involve amino acid residues all over the gene, not just the P1' position right after the start codon. Another hypothesis is that the positively charged lysine helps the protein have stable interactions

with the inner phospholipid membrane [23]. This explanation also doesn't account for the fact that lysine abundance is a genome-wide pattern, and secretory proteins are a minority of proteins.

D) Objective

The objective is to determine the reason behind the prevalence of lysine at the P1' position in *E. coli* and doing so by discriminating against the two hypotheses; one main and one alternative. With the main hypothesis being that lysine is abundant at the P1' position because AAA and AAG are codons for lysine, and this gives the mRNA a reduced secondary structure around the start codon, allowing ribosomes to attach with the mRNA. For alanine, GCA should be a preferred codon to GCG according to this hypothesis; therefore, in general, NNA codons should be greater in number to NNG at P1'. The alternative hypothesis is that the nucleotides are irrelevant, and lysine is abundant in order to prevent the excision of methionine from certain proteins. To test these, codon frequencies of the second amino acid are measured, and nucleotide frequencies of *E. coli* and other bacterial protein coding sequences are also measured.

METHODS

The genomes for all organisms were acquired from the National Center for Biotechnology Information (NCBI). It has a database that stores many different types of data, including genome sequences, which sources the information from numerous studies and compiles them for better accessibility. These publicly available complete genome nucleotide contigs can be downloaded from NCBI in order for further analysis to take place [39]. A total of five genomes were downloaded; their names and NCBI accession code can be found in Table 1; it is a simplified version of a similar table from the previously mentioned paper by Xia in 2026 [13].

RegulonDB, a database that contains information about the transcription initiation of *E. coli* K-12, was used to gather information about the transcription units for the genes. All information about the transcription units was found on the database [40]. A total of 6 transcription units were used and can located in *Supplementary Table 20*.

The Data Analysis in Molecular Biology and Evolution (DAMBE) software is a bioinformatics tool designed for analyzing molecular data, particularly for genomics and phylogenetics. It supports various functions, such as codon usage analysis, motif discovery, Gibbs sampler and other functions [41]. Through these functions, it we analyzed nucleotide distributions and inferred nucleotide distribution patterns and other forms of genome analyses. The codon frequencies were also calculated by DAMBE. All these calculations were done on the whole genome's first 27 nucleotides, including the start codon, since the shortest coding sequence in the annotated genome (ymdG) is only that much long.

The major method used to get an insight into nucleotide distribution patterns was the position weight matrix (PWM), which is a tool in DAMBE [42,43]. The principle behind PWM is in the following equation:

$$PWM_{ij} = log_2 \frac{p_{ij}}{p_i}$$

Where PWM_{ij} is the position specific weight. This method was used to account for background nucleotide frequencies which were computed in DAMBE. The software also adds a pseudo count in order to avoid taking the log of 0 [42]. All statistical analyses were either done by DAMBE automatically or by Microsoft Excel, which is where all graphs were generated.

Species	**Accession code**	**Doubling time (DT)**	**Rank**	**Reference**
Escherichia coli	NC_000913.3	~20—30min	1	[44]
Bacillus subtilis	NC_000964.3	~30—70min	2	[45]
Chlamydia trachomatis	NC_000117.1	~1.8—4.6h	3	[46,47]
Salmonella enterica	NC_003197.2	~20—30min	1	[48]
Mycobacterium tuberculosis	NC_000962.3	~20—30h	4	[49–51]

Table 1. Simplified table of bacteria with doubling times ranked and with accession codes.

RESULTS

A) Codon distribution of lysine at P1'

Hypothesis: Ribosome binding hypothesis (see Introduction)

Alternative hypothesis: Methionine preservation hypothesis (see Introduction)

Predictions: If the hypothesis is true, we would expect the number of AAA to exceed the number of AAG codons, since AAA codons are predicted to produce fewer secondary structures.

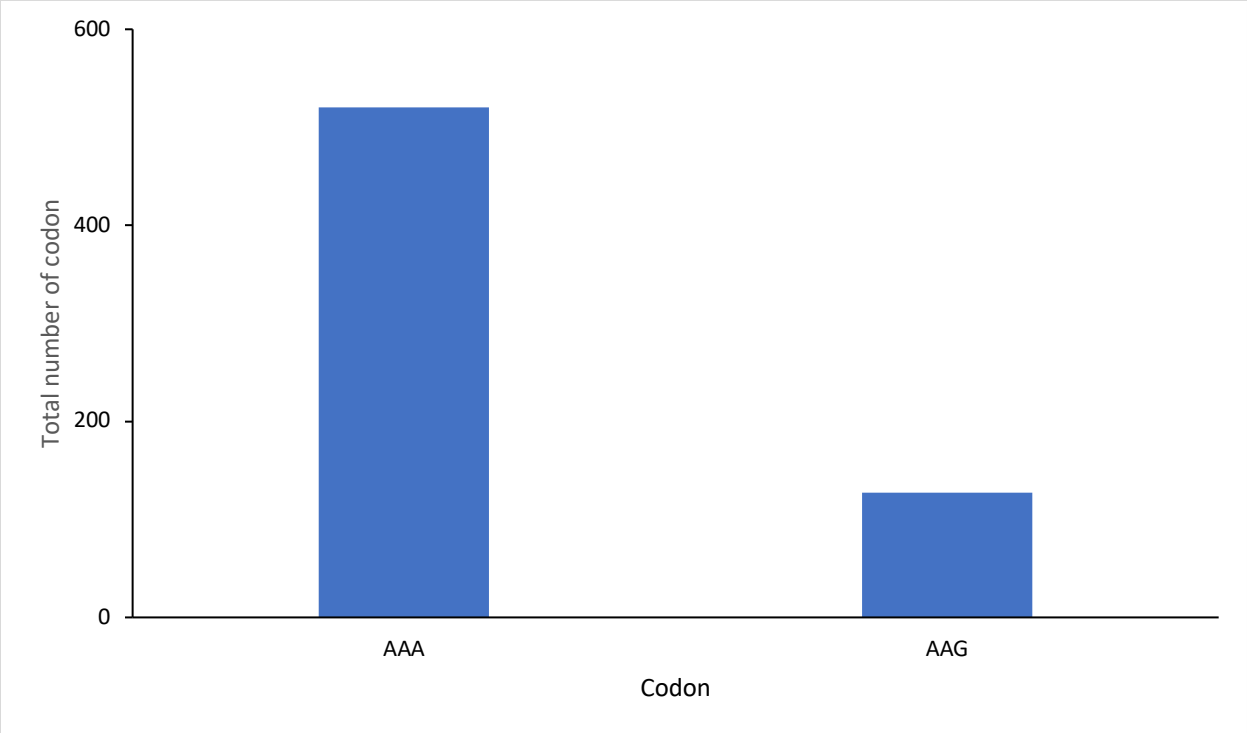


Figure 3. Distribution of lysine codons at the P1' position in *E. coli*

The observed number of AAA codons is significantly higher than the expected 323.5 under the alternate hypothesis (binomial test, $P = 9.5502 \times 10^{-28}$) (see *Supplementary Table 1* for raw data)

Observations: We observed exactly what the primary hypothesis predicted, the AAA codons are more numerous at the P1' position compared to AAG codons. Considering the count is done on the entire genome and the results of the binomial test, the results are considered statistically significant. The binomial test was done with the probability of either codon = 0.5.

B) Codon distribution of alanine at P1'

Hypothesis: Ribosome binding hypothesis (see Introduction)

Alternative hypothesis: Methionine preservation hypothesis (see Introduction)

Predictions: If the hypothesis is true, we would expect the number of GCA to exceed the number of GCG codons, since GCA codons are predicted to produce fewer secondary structures.

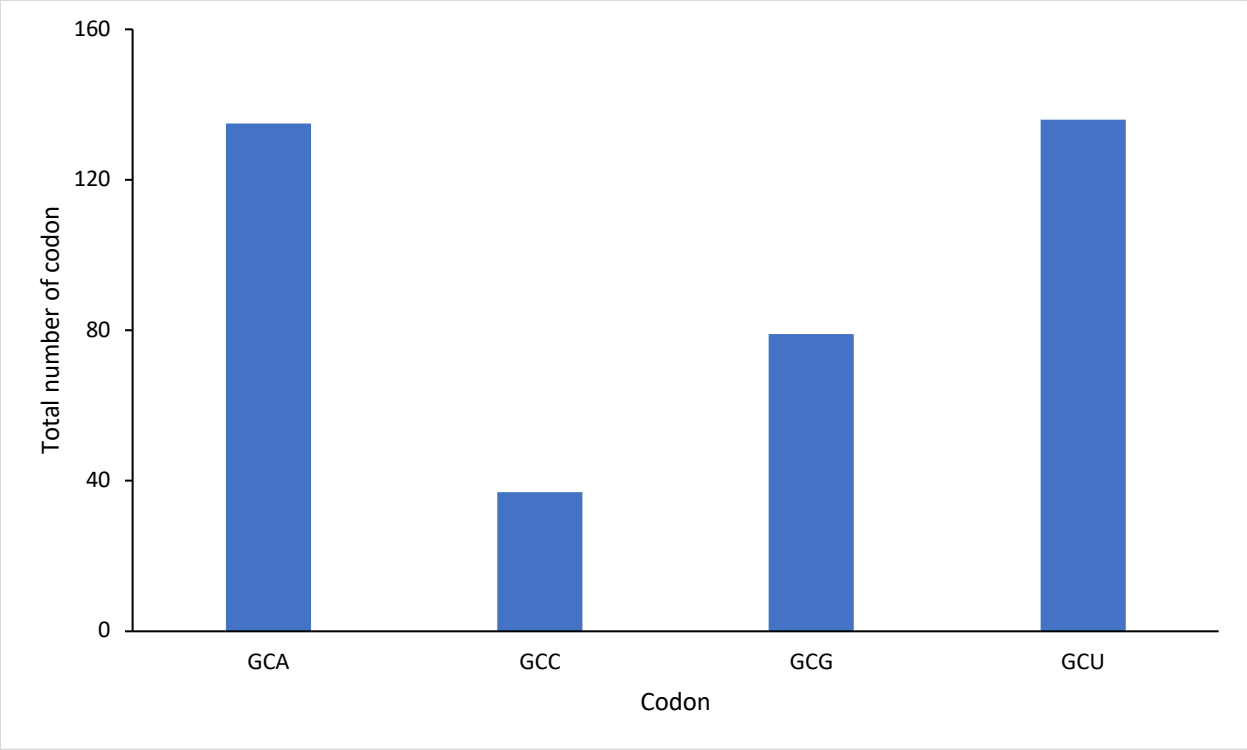


Figure 4. Distribution of alanine codons at second position (P1')

Chi-squared = 6888.75, DF = 3, $p < 0.0001$

(see *Supplementary Table 1* for raw data)

Observations: The results were just as expected under the primary hypothesis. GCU also had an elevated count comparable to GCA, while GCC had the lowest count. A Chi-squared test was conducted with a degree of freedom = 3, giving 6888.75, meaning $p < 0.0001$.

C) Adenine distribution of *E. coli*

Hypothesis: Ribosome binding hypothesis (see Introduction)

Alternative hypothesis: Methionine preservation hypothesis (see Introduction)

Predictions: If the hypothesis is true, we would expect the number of adenines progressively decrease as the reduction of secondary structures becomes less relevant to the transcription start site upstream.

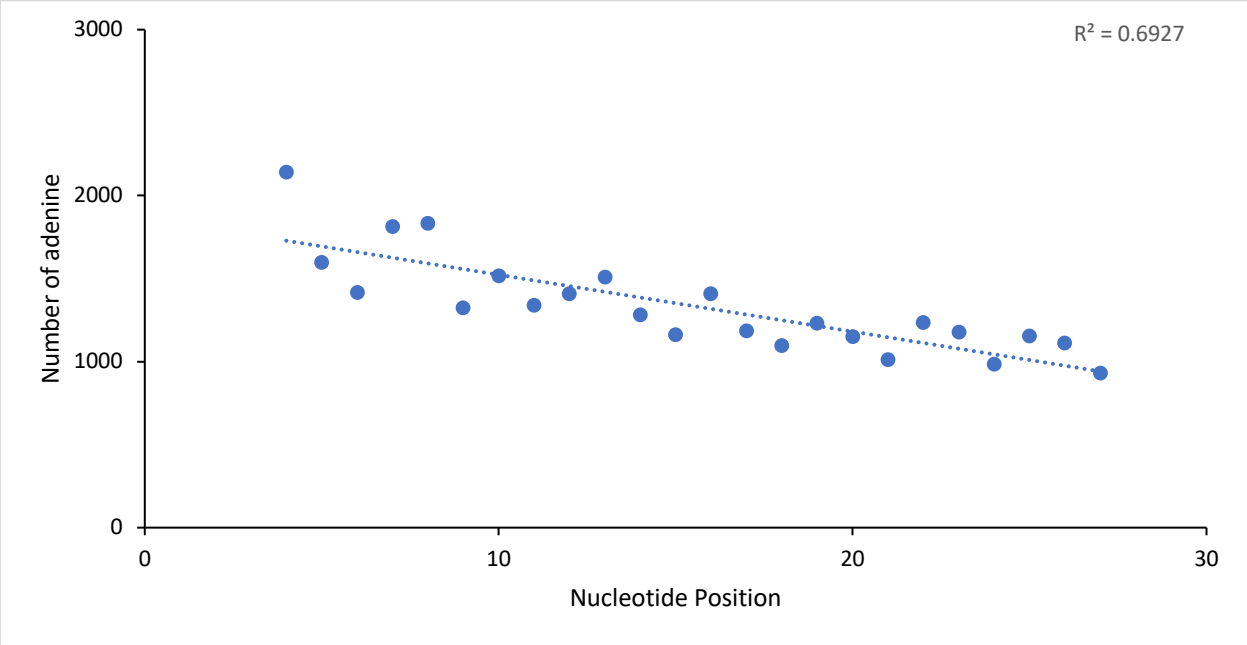


Figure 5. Distribution of adenine in *E. coli* genes across the first 27 nucleotides.

$R^2 = 0.6927$

(see *Supplementary Table 3* for raw data)

Site	A	C	G	U
1	**1.9069**	-8.8822	-1.8025	-3.6749
2	-12.0755	-12.0755	-12.0755	**2.0562**
3	-12.0755	-9.7325	**1.869**	-7.9348
4	**1.0461**	-0.6607	-0.4195	-0.7446
5	**0.6221**	0.2013	-0.7263	-0.4002
6	**0.4497**	-0.3449	-0.7466	0.3907
7	**0.8042**	-0.194	-0.3003	-0.747
8	**0.8208**	-0.3501	-1.0415	0.0541
9	**0.3497**	-0.3588	-0.3808	0.2805
10	**0.548**	-0.0363	-0.466	-0.1889
11	0.3681	-0.1894	-1.0979	**0.5272**
12	**0.4395**	-0.2003	-0.7022	0.2805
13	**0.5394**	-0.102	-0.2616	-0.3129
14	0.3054	-0.133	-1.1164	**0.553**
15	0.1613	-0.1079	-0.606	**0.4283**
16	**0.4415**	0.0478	-0.2719	-0.3164
17	0.192	-0.1123	-1.044	**0.6042**
18	0.0808	-0.1405	-0.3952	**0.3896**
19	**0.2469**	0.0239	0.0064	-0.3373
20	0.1513	-0.1211	-0.8464	**0.5624**
21	-0.0369	0.0661	-0.3984	**0.3245**
22	**0.2516**	0.005	0.0413	-0.3711
23	0.1822	-0.1541	-0.7323	**0.5077**
24	-0.0773	0.0919	-0.34	**0.2919**
25	0.1563	0.0385	**0.1923**	-0.5227
26	0.0978	-0.0962	-0.6266	**0.4821**
27	-0.1572	0.0752	-0.1777	**0.2399**

Table 2. Position Weight Matrix (PWM) of *E. coli* genome for first 27 nucleotides

Highest value PWM score is boldened

Observations: The results were just as expected, the frequency of adenine nucleotides started high and progressively decreased over 27 nucleotides. With an $R^2 = 0.6927$, there is a strong correlation between the trendline and the data points. In the PWM table,

It should be noted that in order to calculate PWM scores, the nucleotide frequencies of the genome were computed in DAMBE prior to the 27-nucleotide restriction (see *Supplementary Table 2*). This allowed for a more accurate reflection of trends across the genome.

D) Different bacterial species' position weight matrix scores for adenine

Hypothesis: Since *E. coli* is a fast-growing bacterium, other fast-growing bacterium as ranked by Table 1 should observe similar adenine nucleotide frequencies, and slower-growing bacteria should show the opposite trend. This assumes the ribosome binding hypothesis is true.

Alternative hypothesis: A specie's doubling time is unrelated to the adenine nucleotide frequencies.

Predictions: All species show decreasing adenine PWM scores, with higher ranked species being over lower ranked species.

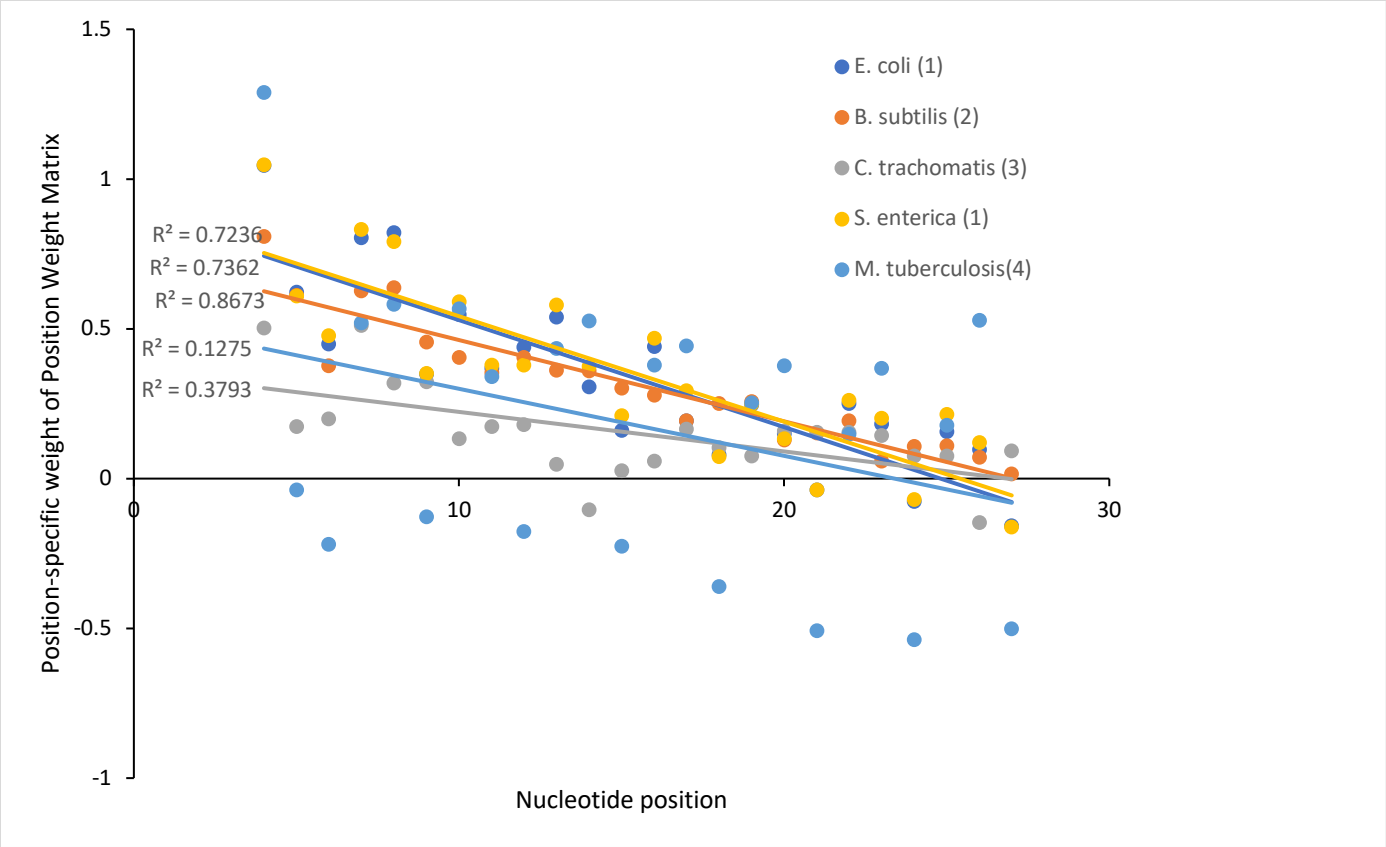


Figure 6. Distribution of adenine in *E. coli, B subtilis, C. trachomatis, S. enterica, M. tuberculosis* genes across the first 27 nucleotides, with their doubling times ranked

$R^2 = 0.7362$, $R^2 = 0.\ 8673$, $R^2 = 0.3793$, $R^2 = 0.7236$, $R^2 = 0.1275$ respectively. The ranks are derived from Table 1. (See *Supplementary Tables 8-19 for all raw data)*

Observations: All bacteria observed follow the same trends as *E. coli* in terms of decreasing adenine frequency over the 27-nucleotide stretch. With the exceptions of *C. trachomatis*, which has data with a moderate level of correlation, and *M. tuberculosis*, which has data with a low level of correlation to the trendline, the other species have a high correlation between data and trendline. Trendlines are roughly arranged in rank, from 1 at the top to 4 at the end, especially near the start codon, however, this is not observed towards the 27th nucleotide mark. For *M. tuberculosis*, the third nucleotide position in the codon seems to diverge from the first two nucleotide positions, the most out of all species.

E) Intragenome *E. coli* position weight matrix scores comparison for adenine

Hypothesis: Assuming the ribosome binding hypothesis is true, the first gene in a polycistronic mRNA would require more adenines near the start codon for the ribosome to bind easier.

Alternative hypothesis: The methionine preservation hypothesis is true.

Predictions: Given the main hypothesis is true, the trendline for the first gene should go progressively down, while the other genes should have a flatter adenine distribution.

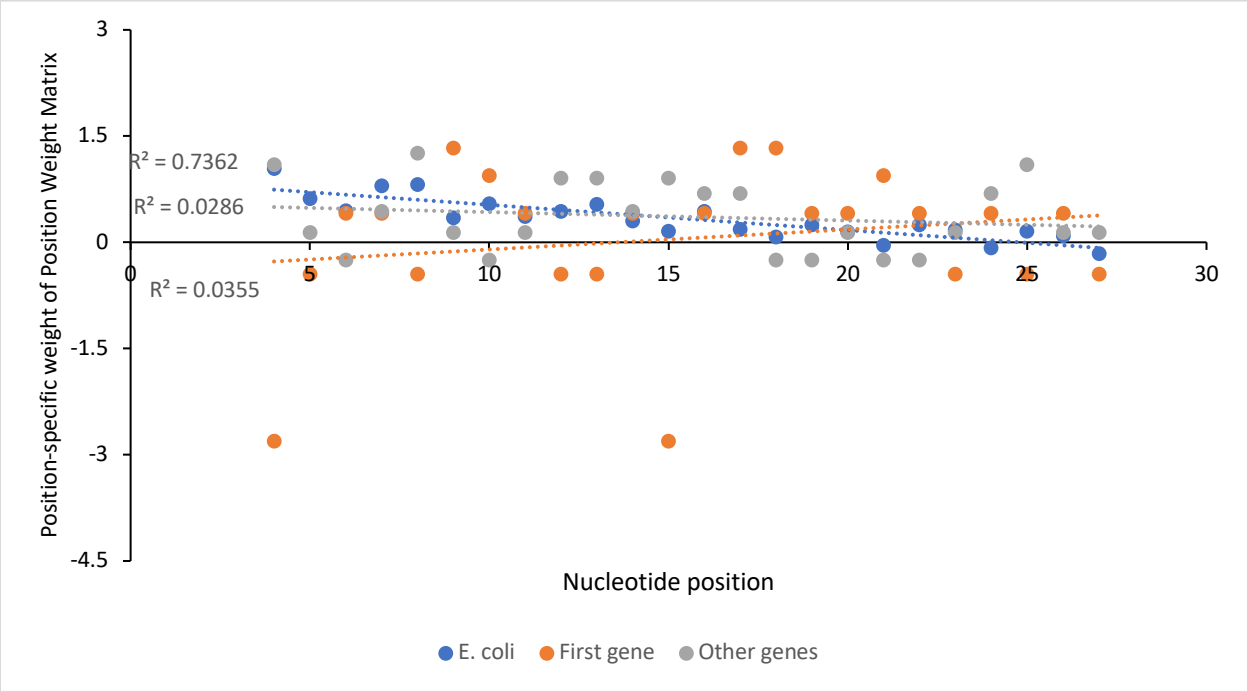


Figure 7. Distribution of adenine in *E. coli* genes, first genes in 6 transcription units, and other genes in transcription units across the first 27 nucleotides

$R^2 = 0.7362$, $R^2 = 0.0355$, $R^2 = 0.0286$ respectively

(See *Supplementary Tables 4-7* for all raw data)

(See *Supplementary Table 20* for genes used from transcription units)

Observations: There are no correlations, as the R^2 values are extremely low. If ignoring that, the trendline is opposite to predictions. No conclusions can be drawn.

DISCUSSION

A) Codon distribution of lysine at P1'

The codon distributions for lysine perfectly support the ribosome binding hypothesis. As we can see in Figure 3, the number of AAA codons was around 4 times higher than AAG, and much higher than the expected 323.5 under equal distribution. The significance was confirmed by a binomial distribution test. With the probability of having either codon n = 0.5, $p = 9.5502 \times 10^{-58}$. However, it is not without doubt, since the probability of having either codon is unlikely to be 0.5. At first glance, it might be different availability of tRNA causing this difference; however, *E. coli* has only 1 tRNA for lysine with a UUU anticodon [52]. Since there is only one tRNA with anticodon, the AAG codon may be less efficient than AAA, favouring the latter in the genome. Nevertheless, these results support the ribosome binding hypothesis.

B) Codon distribution of alanine at P1'

For Figure 4, GCA and GCU are extremely common. GCC seems to be the least common codon at the P1' position for alanine. A chi-squared test confirmed these. On the surface, this may be due to there being simply fewer cytosine nucleotides, as only 20.7% were C nucleotides on a genome wide scale [53], however, this is only the case for human genomes. For *E. coli*, the proportion of each nucleotide in the coding sequences (CDS) is 24.06% for adenine, 24.55% for cytosine, 27.34% for guanine, and 24.04% for thymine (See *Supplementary Table 2*). Evidently, cytosine isn't exceptionally rare in the CDS of *E. coli*, meaning the rarity of cytosines could not have caused the codon to be so low in number. Instead, if we look at tRNA numbers, of which alanine has 2 in *E. coli*, GCU, GCA and GCG are all codons recognized by a single tRNA called

Ala1B accounting for 5.04% of total tRNA, while the Ala2 tRNA which only recognizes GCC accounts for 0.95% of tRNA [52]. This also means that GCA, GCG, and GCU all had the same amount of tRNA, but GCG was less common. This may be due to GCG causing more secondary structure formation. However, GCU and GCA both were about as common at P1'. One possible reason is that GCU doesn't cause as many secondary structure formations. One other reason could be that genes using alanine at the P1' position have no need for more efficient translation because alanine necessarily means guanine and cytosine in the sequence, causing secondary structure formations regardless. In all, since GCA was more common than GCG, the ribosome binding hypothesis is supported.

C) Adenine distribution of *E. coli*

For Figure 5, the adenine nucleotide frequency seems to be steadily decreasing with $R^2 = 0.6927$. The main idea supporting the ribosome binding hypothesis is that a poly(A) stretch is only useful for the first few nucleotides. Once the ribosome has bound, there is no need for a simpler secondary structure for translation elongation. Therefore, there should be a decreasing adenine nucleotide frequency as the sequence is observed more downstream of the start codon. Indeed, we see exactly that. For the position weight matrix of Table 2, 12 out of the 24 nucleotides following AUG favour adenine. For nucleotides 4 through 10, this is an uninterrupted stretch of adenines. There is a very clear preference for adenine closer to the start codon. In addition, adenine would presumably be more easily procured compared to other nucleotides, such as uracil, due to it being present in adenosine triphosphate, one of the most common metabolites. There may also be a limit to the number of adenines that can be added in a poly(A) stretch because in *E. coli*, polyadenylation promotes mRNA degradation [54]. If there

are too many adenosines in a row at the beginning of the mRNA, enzymes responsible for mRNA degradation may degrade the wrong mRNA instead. To end, the results from Figure 5 and Table 2 are perfectly consistent with the ribosome binding hypothesis.

D) Different bacterial species' position weight matrix scores for adenine

In Figure 6, all bacteria observed showed a decreasing frequency of adenines. With the exceptions of *C. trachomatis* and *M. tuberculosis* which had a moderate and low correlation, the other 3 had a high degree of correlation to the trendline. Since all species showed the same trends, it can be additional confirmation that the occurrences of adenine frequencies and the abundance of lysine are not coincidences belonging to *E. coli*. Even more, *M. tuberculosis* having a low R^2 may be due to its GC content. According to computations in DAMBE, *M. tuberculosis* has a GC% of 65.4737%, compared to others where *S. enterica is* 52.3954, *C. trachomatis* 41.5535, *B. subtilis is* 43.1192% *E. coli is* 50.7944% (See *Supplementary Tables 2, 8, 11, 14, 17)*. Due to there being fewer adenines in the first place, there is simply a lower likelihood that any given sequence will have enough adenines to show up on a genome-wide position weight matrix.

Another important pattern was that species that have a faster doubling time also had a higher adenine frequency near the start codon. As the rank of the species doubling time went lower, so did the adenine frequency. Presumably, other species in the same rank as *E. coli* would have a predictable adenine frequency near the AUG codon, while slower growing bacteria would have a predictably lower adenine frequency. For bacteria especially, it makes sense to share translation strategies, since many are competing in speed, one study by Ran et al. in 2012 even said that

selection for translational speed is greater than selection for accuracy [55]. In summary, this supports the ribosome binding hypothesis.

E) Intragenome *E. coli* position weight matrix scores comparison for adenine

The Figure 7 doesn't have statistically relevant trends to speak of; however, there are a few possible reasons why the data did not meet expectations and predictions. Firstly, polycistronic mRNA have a single promoter and terminator [56]. This means that there is only one promoter, which is conserved. Another reason is that all of the transcription units chosen involved some sort of ribosomal protein, which are all highly conserved genes [57]. On the whole, nothing was ascertained from Figure 7, and more tests may be required.

CONCLUSION

To close, in order for the codon frequency conclusions to hold up, more codons at positions other than P1' should be tested for comparison. In addition, more species should have adenine frequencies compared to confirm the patterns in terms of the correlation between doubling time and adenine abundance near the start codon. Furthermore, non-ribosomal proteins belonging to a single transcription unit should have adenine frequencies compared, and more genes should be tested in general. In general, the ribosome binding hypothesis was supported. Most importantly, it is important to examine the conclusions made here in a wet lab experiment for confirmation, as increasing our understanding of translation efficiency would be beneficial in the fields of mRNA vaccines and gene therapeutics, as well as gaining more insight into translation-related diseases.

REFERENCES

[1] Lee CS, Kim D, Hwang CS. Where Does N-Formylmethionine Come from? What for? Where Is It Going? What is the origin of N-formylmethionine in eukaryotic cells? Mol Cells 2022;45:109. https://doi.org/10.14348/MOLCELLS.2021.5040.

[2] Dickerman HW, Steers E, Redfield BG, Weissbach H. Methionyl Soluble Ribonucleic Acid Transformylase: I. PURIFICATION AND PARTIAL CHARACTERIZATION. Journal of Biological Chemistry 1967;242:1522–5. https://doi.org/10.1016/S0021-9258(18)96123-5.

[3] Giglione C, Vallon O, Meinnel T. Control of protein life-span by N-terminal methionine excision. The EMBO Journal 2003 22:1 2003;22:13–23. https://doi.org/10.1093/EMBOJ/CDG007.

[4] Wingfield PT. N-Terminal Methionine Processing. Curr Protoc Protein Sci 2017;88:6.14.1. https://doi.org/10.1002/CPPS.29.

[5] Schechter I, Berger A. On the size of the active site in proteases. I. Papain. Biochem Biophys Res Commun 1967;27:157–62. https://doi.org/10.1016/S0006-291X(67)80055-X.

[6] Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, et al. The complete genome sequence of Escherichia coli K-12. Science 1997;277:1453–62. https://doi.org/10.1126/science.277.5331.1453.

[7] Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, et al. Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. Mol Syst Biol 2006;2:2006.0007. https://doi.org/10.1038/msb4100049.

[8] Jensen KF. The Escherichia coli K-12 “wild types” W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels. J Bacteriol 1993;175:3401–7. https://doi.org/10.1128/JB.175.11.3401-3407.1993.

[9] Bachmann BJ. Pedigrees of some mutant strains of Escherichia coli K-12. Bacteriol Rev 1972;36:525. https://doi.org/10.1128/BR.36.4.525-557.1972.

[10] Shemesh R, Novik A, Cohen Y. Follow the Leader: Preference for Specific Amino Acids Directly Following the Initial Methionine in Proteins of Different Organisms. Genomics Proteomics Bioinformatics 2010;8:180–9. https://doi.org/10.1016/S1672-0229(10)60020-4.

[11] Bivona L, Zou Z, Stutzman N, Sun PD. Influence of the Second Amino Acid on Recombinant Protein Expression. Protein Expr Purif 2010;74:248. https://doi.org/10.1016/J.PEP.2010.06.005.

[12] Gibson B, Wilson DJ, Feil E, Eyre-Walker A. The distribution of bacterial doubling times in the wild. Proceedings of the Royal Society B: Biological Sciences 2018;285:20180789. https://doi.org/10.1098/rspb.2018.0789.

[13] Xia X. Evolution of Translational Machinery in Fast- and Slow-Growing Bacteria. Microorganisms 2026, Vol 14, Page 377 2026;14:377. https://doi.org/10.3390/MICROORGANISMS14020377.

[14] Li GW. How do bacteria tune translation efficiency? Curr Opin Microbiol 2015;24:66. https://doi.org/10.1016/J.MIB.2015.01.001.

[15] Ito K, Chiba S, Pogliano K. Divergent stalling sequences sense and control cellular physiology. Biochem Biophys Res Commun 2010;393:1–5. https://doi.org/10.1016/J.BBRC.2010.01.073.

[16] Baranov P V., Gesteland RF, Atkins JF. Recoding: translational bifurcations in gene expression. Gene 2002;286:187–201. https://doi.org/10.1016/S0378-1119(02)00423-7.

[17] Frottin F, Martinez A, Peynot P, Mitra S, Holz RC, Giglione C, et al. The proteomics of N-terminal methionine cleavage. Molecular and Cellular Proteomics 2006;5:2336–49. https://doi.org/10.1074/mcp.M600225-MCP200.

[18] Khalouei S, Yao X, Mennigen J, Carullo M, Ma P, Song Z, et al. Bioinformatic approach to identify penultimate amino acids efficient for N-terminal methionine excision. 2007 1st International Conference on Bioinformatics and Biomedical Engineering, ICBBE 2007:386–9. https://doi.org/10.1109/ICBBE.2007.102.

[19] Hirel PH, Schmitter JM, Dessen P, Fayat G, Blanquet S. Extent of N-terminal methionine excision from Escherichia coli proteins is governed by the side-chain length of the penultimate amino acid. Proc Natl Acad Sci U S A 1989;86:8247–51. https://doi.org/10.1073/PNAS.86.21.8247.

[20] Wingfield PT. N-Terminal Methionine Processing. Curr Protoc Protein Sci 2017;88:6.14.1. https://doi.org/10.1002/CPPS.29.

[21] Giglione C, Boularot A, Meinnel T. Protein N-terminal methionine excision. Cellular and Molecular Life Sciences 2004;61:1455–74. https://doi.org/10.1007/S00018-004-3466-8/METRICS.

[22] Chang YH. Impact of Protein Nα-Modifications on Cellular Functions and Human Health. Life 2023, Vol 13, Page 1613 2023;13:1613. https://doi.org/10.3390/LIFE13071613.

[23] Zalucki YM, Power PM, Jennings MP. Selection for efficient translation initiation biases codon usage at second amino acid position in secretory proteins. Nucleic Acids Res 2007;35:5748. https://doi.org/10.1093/NAR/GKM577.

[24] Xiao Q, Zhang F, Nacev BA, Liu JO, Pei D. Protein N-Terminal Processing: Substrate Specificity of Escherichia coli and Human Methionine Aminopeptidases. Biochemistry 2010;49:5588. https://doi.org/10.1021/BI1005464.

[25] Huson DH, Scornavacca C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst Biol 2012;61:1061–7. https://doi.org/10.1093/SYSBIO/SYS062.

[26] Lu JP, Chai SC, Ye QZ. Catalysis and Inhibition of Mycobacterium tuberculosis Methionine Aminopeptidase. J Med Chem 2009;53:1329–37. https://doi.org/10.1021/JM901624N.

[27] Kanudia P, Mittal M, Kumaran S, Chakraborti PK. Amino-terminal extension present in the methionine aminopeptidase type 1c of Mycobacterium tuberculosis is indispensible for its activity. BMC Biochem 2011;12:35. https://doi.org/10.1186/1471-2091-12-35.

[28] Fricke M, Gerst R, Ibrahim B, Niepmann M, Marz M. Global importance of RNA secondary structures in protein-coding sequences. Bioinformatics 2018;35:579. https://doi.org/10.1093/bioinformatics/bty678.

[29] Ma Q, Zhang X, Yang J, Li H, Hao Y, Feng X. Optimization of the 5′ untranslated region of mRNA vaccines. Sci Rep 2024;14:19845. https://doi.org/10.1038/s41598-024-70792-x.

[30] Karlin S, Doerfler W, Cardon LR. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J Virol 1994;68:2889–97. https://doi.org/10.1128/jvi.68.5.2889-2897.1994.

[31] Hall MN, Gabay J, Débarbouillé M, Schwartz M. A role for mRNA secondary structure in the control of translation initiation. Nature 1982 295:5850 1982;295:616–8. https://doi.org/10.1038/295616a0.

[32] Georgakopoulos-Soares I, Parada GE, Hemberg M. Secondary structures in RNA synthesis, splicing and translation. Comput Struct Biotechnol J 2022;20:2871. https://doi.org/10.1016/J.CSBJ.2022.05.041.

[33] Fricke M, Gerst R, Ibrahim B, Niepmann M, Marz M. Global importance of RNA secondary structures in protein-coding sequences. Bioinformatics 2018;35:579. https://doi.org/10.1093/BIOINFORMATICS/BTY678.

[34] Pelletier J, Sonenberg N. The involvement of mRNA secondary structure in protein synthesis. Https://DoiOrg/101139/O87-074 2011;65:576–81. https://doi.org/10.1139/O87-074.

[35] Ali I, Conrad RJ, Verdin E, Ott M. Lysine Acetylation Goes Global: From Epigenetics to Metabolism and Therapeutics. Chem Rev 2018;118:1216. https://doi.org/10.1021/acs.chemrev.7b00181.

[36] Shimko JC, North JA, Bruns AN, Poirier MG, Ottesen JJ. Preparation of Fully Synthetic Histone H3 Reveals That Acetyl-Lysine 56 Facilitates Protein Binding Within Nucleosomes. J Mol Biol 2011;408:187–204. https://doi.org/10.1016/j.jmb.2011.01.003.

[37] Neumann H, Hancock SM, Buning R, Routh A, Chapman L, Somers J, et al. A Method for Genetically Installing Site-Specific Acetylation in Recombinant Histones Defines the Effects of H3 K56 Acetylation. Mol Cell 2009;36:153–63. https://doi.org/10.1016/j.molcel.2009.07.027.

[38] Jing Y, Li X, Liu Z, Li XD. Roles of Negatively Charged Histone Lysine Acylations in Regulating Nucleosome Structure and Dynamics. Front Mol Biosci 2022;9:899013. https://doi.org/10.3389/fmolb.2022.899013.

[39] Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Connor R, et al. Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Res 2024;53:D20. https://doi.org/10.1093/NAR/GKAE979.

[40] Salgado H, Gama-Castro S, Lara P, Mejia-Almonte C, Alarcón-Carranza G, López-Almazo AG, et al. RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12. Nucleic Acids Res 2024;52:D255–64. https://doi.org/10.1093/NAR/GKAD1072.

[41] Xia X. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol 2013;30:1720–8. https://doi.org/10.1093/molbev/mst064.

[42] Xia X. Position weight matrix and Perceptron. Bioinformatics and the cell: modern computational approaches in genomics, proteomics, and transcriptomics, vol. Cham. 2nd edition, Springer; 2018, p. 77–98.

[43] Xia X. Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction. Scientifica (Cairo) 2012;2012.

[44] Sezonov G, Joseleau-Petit D, D'Ari R. Escherichia coli physiology in Luria-Bertani broth. J Bacteriol 2007;189:8746–9. https://doi.org/10.1128/JB.01368-07;WEBSITE:WEBSITE:ASMJ;REQUESTEDJOURNAL:JOURNAL:JB;JOURNAL:JOURNAL:JB;WGROUP:STRING:PUBLICATION.

[45] Sharpe ME, Hauser PM, Sharpe RG, Errington J. Bacillus subtilis cell cycle as studied by fluorescence microscopy: Constancy of cell length at initiation of DNA replication and

evidence for active nucleoid partitioning. J Bacteriol 1998;180:547–55. https://doi.org/10.1128/JB.180.3.547-555.1998;ISSUE:ISSUE:DOI.

[46] Hayward RJ, Marsh JW, Humphrys MS, Huston WM, Myers GSA. Early Transcriptional Landscapes of Chlamydia trachomatis-Infected Epithelial Cells at Single Cell Resolution. Front Cell Infect Microbiol 2019;9:489901. https://doi.org/10.3389/FCIMB.2019.00392/TEXT.

[47] Lee JK, Enciso GA, Boassa D, Chander CN, Lou TH, Pairawan SS, et al. Replication-dependent size reduction precedes differentiation in Chlamydia trachomatis. Nature Communications 2017 9:1 2018;9:45-. https://doi.org/10.1038/s41467-017-02432-0.

[48] Silva RR, Moraes CA, Bessan J, Vanetti MCD. Validation of a predictive model describing growth of Salmonella in enteral feeds. Brazilian Journal of Microbiology 2009;40:149–54. https://doi.org/10.1590/S1517-83822009000100026.

[49] Gengenbacher M, Kaufmann SHE. Mycobacterium tuberculosis: success through dormancy. FEMS Microbiol Rev 2012;36:514–32. https://doi.org/10.1111/J.1574-6976.2012.00331.X.

[50] Zhu M, Dai X. On the intrinsic constraint of bacterial growth rate: M. tuberculosis's view of the protein translation capacity. Crit Rev Microbiol 2018;44:455–64. https://doi.org/10.1080/1040841X.2018.1425672.

[51] Cole ST. Comparative and functional genomics of the Mycobacterium tuberculosis complex. Microbiology (N Y) 2002;148:2919–28. https://doi.org/10.1099/00221287-148-10-2919/CITE/REFWORKS.

[52] Dong H, Nilsson L, Kurland CG. Co-variation of tRNA Abundance and Codon Usage inEscherichia coliat Different Growth Rates. J Mol Biol 1996;260:649–63. https://doi.org/10.1006/JMBI.1996.0428.

[53] Yamagishi MEB, Shimabukuro AI. Nucleotide Frequencies in Human Genome and Fibonacci Numbers. Bulletin of Mathematical Biology 2007 70:3 2007;70:643–53. https://doi.org/10.1007/S11538-007-9261-6.

[54] Régnier P, Marujo PE. Polyadenylation and Degradation of RNA in Prokaryotes 2013.

[55] Ran W, Higgs PG. Contributions of Speed and Accuracy to Translational Selection in Bacteria. PLoS One 2012;7:e51652. https://doi.org/10.1371/JOURNAL.PONE.0051652.

[56] Ghavidel AA, Jajarmi V, Bandehpour M, Kazemi B. Polycistronic Expression of Multi-Subunit Complexes in the Eukaryotic Environment: A Narrative Review. Iran J Parasitol 2022;17:286. https://doi.org/10.18502/IJPA.V17I3.10618.

[57] Alksne LE, Anthony RA, Liebman SW, Warner JR. An accuracy center in the ribosome conserved over 2 billion years. Proc Natl Acad Sci U S A 1993;90:9538–41. https://doi.org/10.1073/PNAS.90.20.9538;WGROUP:STRING:PUBLICATION.

APPENDIX

Supplementary Figure 1. Codon table 11 from DAMBE

Supplementary Table 1. Observed di-codon frequencies in CDS for *E. coli*

Supplementary Table 2. Background frequency of nucleotides in *E. coli*

Supplementary Table 3. Site-specific frequencies of nucleotides in *E. coli* for first 27 nucleotides

Supplementary Table 4. Site-specific frequencies of nucleotides in *E. coli* for first 27 nucleotides of first gene in transcription unit

Supplementary Table 5. PWM of nucleotides in *E. coli* for first 27 nucleotides of first gene in transcription unit

Supplementary Table 6. Site-specific frequencies of nucleotides in *E. coli* for first 27 nucleotides of other genes in transcription unit

Supplementary Table 7. PWM of nucleotides in *E. coli* for first 27 nucleotides of other genes in transcription unit

Supplementary Table 8. Background frequency of nucleotides in CDS for *B. subtilis*

Supplementary Table 9. Site-specific frequencies of nucleotides in *B. subtilis* for first 27 nucleotides

Supplementary Table 10. PWM of nucleotides in *B. subtilis* for first 27 nucleotides of other genes in transcription unit

Supplementary Table 11. Background frequency of nucleotides in CDS for *C. trachomatis*

Supplementary Table 12. Site-specific frequencies of nucleotides in *C. trachomatis* for first 27 nucleotides

Supplementary Table 13. PWM of nucleotides in *C. trachomatis* for first 27 nucleotides of other genes in transcription unit

Supplementary Table 14. Background frequency of nucleotides in CDS for *S. enterica*

Supplementary Table 15. Site-specific frequencies of nucleotides in *S. enterica* for first 27 nucleotides

Supplementary Table 16. PWM of nucleotides in *S. enterica* for first 27 nucleotides of other genes in transcription unit

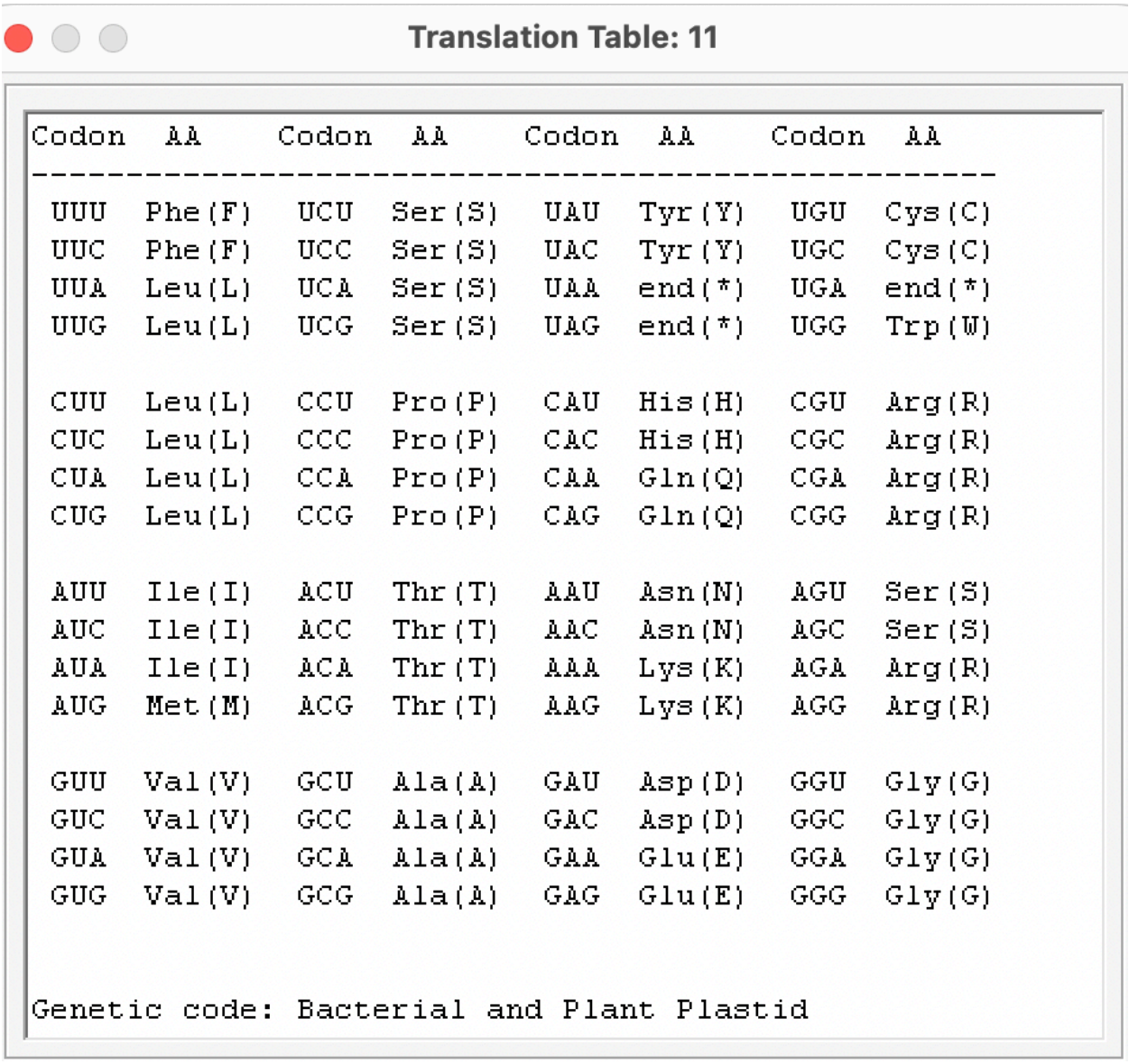
Supplementary Table 17. Background frequency of nucleotides in CDS for *M. tuberculosis*

Supplementary Table 18. Site-specific frequencies of nucleotides in *M. tuberculosis* for first 27 nucleotides

Supplementary Table 19. PWM of nucleotides in *M. tuberculosis* for first 27 nucleotides of other genes in transcription unit

Supplementary Table 20. Genes separated by position in transcription unit

Supplementary Figure 1. Codon table 11 from DAMBE



Supplementary Table 1. Observed di-codon frequencies in *E. coli*

Sequences lengths were restricted to 6 nucleotides, the leftmost column represents the first codon, and the top row represents the second codon.

	AAA	AAG	GCA	GCC	GCG	GCU
AUG	466	113	122	34	64	123
AUU	1	0	0	0	0	0
GUG	47	12	10	2	11	11
UUG	6	2	3	1	4	2

Supplementary Table 2. Background frequency of nucleotides in CDS for *E. coli*

Nuc	A	C	G	T
ObsN	968755	988430	1100861	967828
Prop	0.2406	0.2455	0.2734	0.2404

GC% = 50.7944
GT% = 51.3286

Supplementary Table 3. Site-specific frequencies of nucleotides in *E. coli* for first 27 nucleotides

Site	A	C	G	U	X2	p
1	3894	2	338	81	7208.163	0.0000000
2	0	0	0	4315	11443.231	0.0000000
3	0	1	4310	4	15739.983	0.0000000
4	2144	670	882	619	815.518	0.0000003
5	1598	1218	713	786	368.322	0.0000002
6	1418	834	703	1360	88.531	0.0000001
7	1813	926	958	618	447.426	0.0000002
8	1834	831	573	1077	336.621	0.0000002
9	1323	826	906	1260	8.596	0.0351770
10	1518	1033	854	910	122.367	0.0000001
11	1340	929	551	1495	238.401	0.0000002
12	1408	922	725	1260	55.791	0.0000001
13	1509	987	984	835	143.069	0.0000001
14	1283	966	544	1522	266.722	0.0000002
15	1161	983	775	1396	98.245	0.0000001
16	1410	1095	977	833	164.840	0.0000001
17	1186	980	572	1577	295.834	0.0000002
18	1098	961	897	1359	76.455	0.0000001
19	1232	1077	1185	821	235.545	0.0000002
20	1153	974	656	1532	217.412	0.0000002
21	1012	1109	895	1299	150.818	0.0000001
22	1236	1063	1214	802	257.789	0.0000002
23	1178	952	710	1475	147.429	0.0000001
24	984	1129	932	1270	169.821	0.0000001
25	1157	1088	1348	722	444.543	0.0000003
26	1111	991	764	1449	140.240	0.0000001
27	931	1116	1043	1225	201.908	0.0000001

Supplementary Table 4. Site-specific frequencies of nucleotides in *E. coli* for first 27 nucleotides of first gene in transcription unit

Site	A	C	G	U	X2	p
1	6	0	0	0	13.059	0.0045111
2	0	0	0	6	18.300	0.0003814
3	0	0	6	0	22.588	0.0000492
4	0	1	3	2	4.577	0.2055424
5	1	4	0	1	6.880	0.0758196
6	2	3	0	1	3.360	0.3393494
7	2	1	2	1	0.699	0.8734750
8	1	2	1	2	0.942	0.8151738
9	4	0	1	1	3.940	0.2680488
10	3	1	2	0	2.671	0.4451942
11	2	0	0	4	6.918	0.0745695
12	1	1	0	4	6.059	0.1087681
13	1	3	0	2	3.797	0.2842377
14	2	0	2	2	1.994	0.5736285
15	0	2	2	2	2.795	0.4242594
16	2	3	0	1	3.360	0.3393494
17	4	1	0	1	3.875	0.2752461
18	4	1	0	1	3.875	0.2752461
19	2	0	4	0	8.824	0.0317316
20	2	1	1	2	0.341	0.9520446
21	3	1	0	2	2.194	0.5330452
22	2	3	1	0	3.479	0.3234526
23	1	2	3	0	4.595	0.2039380
24	2	1	0	3	2.922	0.4037486
25	1	2	3	0	4.595	0.2039380
26	2	1	3	0	3.994	0.2620656
27	1	3	0	2	3.797	0.2842377

Supplementary Table 5. PWM of nucleotides in *E. coli* for first 27 nucleotides of first gene in transcription unit

Site	A	C	G	U
1	1.8896	-2.8074	-2.8074	-2.8074
2	-2.8074	-2.8074	-2.8074	1.8908
3	-2.8074	-2.8074	1.7128	-2.8074
4	-2.8074	-0.4644	0.7743	0.4129
5	-0.4410	1.3048	-2.8074	-0.4401
6	0.4118	0.9173	-2.8074	-0.4401
7	0.4118	-0.4644	0.2484	-0.4401
8	-0.4410	0.3859	-0.5878	0.4129
9	1.3322	-2.8074	-0.5878	-0.4401
10	0.9442	-0.4644	0.2484	-2.8074
11	0.4118	-2.8074	-2.8074	1.3333
12	-0.4410	-0.4644	-2.8074	1.3333
13	-0.4410	0.9173	-2.8074	0.4129
14	0.4118	-2.8074	0.2484	0.4129
15	-2.8074	0.3859	0.2484	0.4129
16	0.4118	0.9173	-2.8074	-0.4401
17	1.3322	-0.4644	-2.8074	-0.4401
18	1.3322	-0.4644	-2.8074	-0.4401
19	0.4118	-2.8074	1.1589	-2.8074
20	0.4118	-0.4644	-0.5878	0.4129
21	0.9442	-0.4644	-2.8074	0.4129
22	0.4118	0.9173	-0.5878	-2.8074
23	-0.4410	0.3859	0.7743	-2.8074
24	0.4118	-0.4644	-2.8074	0.9453
25	-0.4410	0.3859	0.7743	-2.8074
26	0.4118	-0.4644	0.7743	-2.8074
27	-0.4410	0.9173	-2.8074	0.4129

Supplementary Table 6. Site-specific frequencies of nucleotides in *E. coli* for first 27 nucleotides of other genes in transcription unit

Site	A	C	G	U	X2	p
1	13	0	1	1	19.854	0.0001820
2	0	0	0	15	38.761	0.0000001
3	0	0	15	0	54.034	0.0000000
4	8	1	4	2	4.238	0.2369237
5	4	7	3	1	9.882	0.0195943
6	3	2	4	6	1.859	0.6021689
7	5	5	2	3	2.960	0.3978421
8	9	0	2	4	6.494	0.0899038
9	4	4	1	6	3.241	0.3559085
10	3	3	5	4	1.743	0.6274020
11	4	2	2	7	2.704	0.4394658
12	7	2	1	5	2.749	0.4320038
13	7	2	4	2	2.333	0.5061643
14	5	4	1	5	2.440	0.4862386
15	7	3	1	4	2.500	0.4753398
16	6	3	5	1	3.640	0.3030289
17	6	3	3	3	0.643	0.8866117
18	3	3	3	6	1.613	0.6565054
19	3	2	9	1	13.439	0.0037766
20	4	3	3	5	0.405	0.9391101
21	3	3	5	4	1.743	0.6274020
22	3	4	5	3	2.732	0.4347372
23	4	2	2	7	2.704	0.4394658
24	6	4	3	2	2.110	0.5499138
25	8	2	3	2	3.231	0.3574016
26	4	4	0	7	6.041	0.1096520
27	4	3	1	7	3.685	0.2974957

Supplementary Table 7. PWM of nucleotides in *E. coli* for first 27 nucleotides of other genes in transcription unit

Site	A	C	G	U
1	1.7822	-4.0000	-1.7804	-1.6327
2	-4.0000	-4.0000	-4.0000	1.9863
3	-4.0000	-4.0000	1.8039	-4.0000
4	1.0980	-1.6571	-0.0337	-0.7798
5	0.1396	0.8833	-0.4183	-1.6327
6	-0.2484	-0.8068	-0.0337	0.6981
7	0.4450	0.4173	-0.9442	-0.2473
8	1.2633	-4.0000	-0.9442	0.1407
9	0.1396	0.1121	-1.7804	0.6981
10	-0.2484	-0.2754	0.2696	0.1407
11	0.1396	-0.8068	-0.9442	0.9126
12	0.9114	-0.8068	-1.7804	0.4462
13	0.9114	-0.8068	-0.0337	-0.7798
14	0.4450	0.1121	-1.7804	0.4462
15	0.9114	-0.2754	-1.7804	0.1407
16	0.6970	-0.2754	0.2696	-1.6327
17	0.6970	-0.2754	-0.4183	-0.2473
18	-0.2484	-0.2754	-0.4183	0.6981
19	-0.2484	-0.8068	1.0840	-1.6327
20	0.1396	-0.2754	-0.4183	0.4462
21	-0.2484	-0.2754	0.2696	0.1407
22	-0.2484	0.1121	0.2696	-0.2473
23	0.1396	-0.8068	-0.9442	0.9126
24	0.6970	0.1121	-0.4183	-0.7798
25	1.0980	-0.8068	-0.4183	-0.7798
26	0.1396	0.1121	-4.0000	0.9126
27	0.1396	-0.2754	-1.7804	0.9126

Supplementary Table 8. Background frequency of nucleotides in CDS for *B. subtilis*

Summary output:

Nuc	A	C	G	T
ObsN	1113418	747151	890812	953112
Prop	0.3006	0.2017	0.2405	0.2573

GC% = 43.1192
GT% = 49.7127

Supplementary Table 9. Site-specific frequencies of nucleotides in *B. subtilis* for first 27 nucleotides

Site	A	C	G	U	X2	p
1	3292	5	387	553	3215.417	0.0000000
2	0	0	0	4237	11051.080	0.0000000
3	0	1	4228	8	15848.431	0.0000000
4	2232	448	1015	542	718.155	0.0000003
5	1946	879	603	809	410.067	0.0000002
6	1653	565	765	1254	37.481	0.0000001
7	1967	724	906	640	373.993	0.0000002
8	1980	642	520	1095	283.196	0.0000002
9	1747	568	795	1127	45.084	0.0000001
10	1685	781	947	824	157.513	0.0000001
11	1626	708	514	1389	212.304	0.0000001
12	1687	594	742	1214	41.435	0.0000001
13	1636	770	1009	822	158.414	0.0000001
14	1634	665	515	1423	218.037	0.0000001
15	1571	586	807	1273	19.596	0.0002059
16	1546	728	1100	863	147.381	0.0000001
17	1456	701	529	1551	280.255	0.0000002
18	1516	681	811	1229	14.201	0.0026433
19	1522	729	1130	856	166.894	0.0000001
20	1393	763	562	1519	264.847	0.0000002
21	1415	794	799	1229	65.292	0.0000001
22	1457	732	1192	856	210.116	0.0000001
23	1326	774	581	1556	294.918	0.0000002
24	1372	781	833	1251	63.920	0.0000001
25	1375	799	1236	827	300.960	0.0000002
26	1339	785	638	1475	213.595	0.0000001
27	1287	801	866	1283	99.730	0.0000001

Supplementary Table 10. PWM of nucleotides in *B. subtilis* for first 27 nucleotides of other genes in transcription unit

Site	A	C	G	U
1	1.3698	-7.3605	-1.3962	-0.9789
2	-12.0492	-12.0492	-12.0492	1.9582
3	-12.0492	-9.4744	2.0526	-7.0450
4	0.8092	-0.9314	-0.0057	-1.0079
5	0.6114	0.0406	-0.7567	-0.4302
6	0.3760	-0.5968	-0.4135	0.2019
7	0.6269	-0.2392	-0.1695	-0.7682
8	0.6364	-0.4126	-0.9702	0.0064
9	0.4558	-0.5892	-0.3580	0.0479
10	0.4037	-0.1299	-0.1057	-0.4037
11	0.3523	-0.2714	-0.9870	0.3494
12	0.4054	-0.5246	-0.4575	0.1552
13	0.3611	-0.1504	-0.0142	-0.4072
14	0.3594	-0.3618	-0.9842	0.3843
15	0.3027	-0.5442	-0.3364	0.2236
16	0.2795	-0.2313	0.1103	-0.3370
17	0.1930	-0.2858	-0.9455	0.5085
18	0.2513	-0.3275	-0.3293	0.1729
19	0.2570	-0.2293	0.1491	-0.3488
20	0.1292	-0.1635	-0.8582	0.4785
21	0.1518	-0.1061	-0.3508	0.1729
22	0.1940	-0.2234	0.2262	-0.3488
23	0.0581	-0.1429	-0.8103	0.5132
24	0.1073	-0.1299	-0.2907	0.1985
25	0.1104	-0.0970	0.2785	-0.3985
26	0.0722	-0.1225	-0.6753	0.4361
27	0.0150	-0.0934	-0.2347	0.2349

Supplementary Table 11. Background frequency of nucleotides in CDS for *C. trachomatis*

Summary output:

Nuc	A	C	G	T
ObsN	266455	187393	202705	279771
Prop	0.2846	0.2001	0.2165	0.2988

GC% = 41.5535
GT% = 51.6544

Supplementary Table 12. Site-specific frequencies of nucleotides in *C. trachomatis* for first 27 nucleotides

Site	A	C	G	U	X2	p
1	784	0	71	32	1344.766	0.0000004
2	0	0	0	887	1979.770	0.0000000
3	0	0	887	0	3511.088	0.0000000
4	358	157	213	159	77.221	0.0000001
5	285	221	163	218	42.008	0.0000000
6	290	155	149	293	6.654	0.0837994
7	360	155	192	180	56.712	0.0000000
8	315	183	142	247	20.240	0.0001514
9	316	134	157	280	10.244	0.0166032
10	277	188	191	231	15.098	0.0017348
11	285	202	112	288	40.567	0.0000000
12	286	149	141	311	13.229	0.0041669
13	261	185	210	231	19.627	0.0002029
14	235	211	132	309	44.570	0.0000000
15	257	154	127	349	37.102	0.0000001
16	263	177	223	224	24.465	0.0000200
17	283	172	130	302	18.214	0.0003973
18	271	129	137	350	34.999	0.0000002
19	266	184	206	231	17.363	0.0005950
20	282	192	131	282	22.286	0.0000569
21	281	155	128	323	23.080	0.0000389
22	281	170	220	216	23.509	0.0000316
23	279	187	125	296	24.876	0.0000164
24	266	159	136	326	20.757	0.0001183
25	266	189	219	213	31.218	0.0000008
26	228	218	149	292	41.913	0.0000000
27	269	139	139	340	26.451	0.0000077

Supplementary Table 13. PWM of nucleotides in *C. trachomatis* for first 27 nucleotides of other genes in transcription unit

Site	A	C	G	U
1	1.6338	-9.7944	-1.4327	-3.0383
2	-9.7944	-9.7944	-9.7944	1.7416
3	-9.7944	-9.7944	2.2063	-9.7944
4	0.5035	-0.1768	0.1493	-0.7361
5	0.1748	0.3160	-0.2362	-0.2815
6	0.1999	-0.1952	-0.3656	0.1446
7	0.5116	-0.1952	-0.0003	-0.5574
8	0.3191	0.0441	-0.4349	-0.1016
9	0.3237	-0.4050	-0.2902	0.0792
10	0.1338	0.0829	-0.0078	-0.1981
11	0.1748	0.1864	-0.7767	0.1198
12	0.1799	-0.2521	-0.4451	0.2305
13	0.0481	0.0597	0.1289	-0.1981
14	-0.1032	0.2493	-0.5401	0.2212
15	0.0258	-0.2045	-0.5957	0.3966
16	0.0591	-0.0040	0.2154	-0.2424
17	0.1647	-0.0453	-0.5621	0.1882
18	0.1022	-0.4597	-0.4865	0.4008
19	0.0754	0.0519	0.1012	-0.1981
20	0.1596	0.1133	-0.5510	0.0894
21	0.1545	-0.1952	-0.5844	0.2851
22	0.1545	-0.0621	0.1959	-0.2948
23	0.1442	0.0752	-0.6186	0.1592
24	0.0754	-0.1585	-0.4971	0.2984
25	0.0754	0.0906	0.1894	-0.3149
26	-0.1467	0.2963	-0.3656	0.1396
27	0.0916	-0.3522	-0.4657	0.3590

Supplementary Table 14. Background frequency of nucleotides in CDS for *S. enterica*

Summary output:

Nuc	A	C	G	T
ObsN	984369	1063605	1187234	985492
Prop	0.2332	0.2520	0.2813	0.2335

GC% = 52.3954
GT% = 51.4698

Supplementary Table 15. Site-specific frequencies of nucleotides in *S. enterica* for first 27 nucleotides

Site	A	C	G	U	X2	p
1	3965	12	367	108	7372.893	0.0000000
2	0	0	0	4452	11935.439	0.0000000
3	0	0	4452	0	15575.252	0.0000000
4	2147	748	937	620	794.397	0.0000003
5	1584	1254	719	895	330.827	0.0000002
6	1446	822	850	1334	47.862	0.0000001
7	1847	927	1007	671	426.878	0.0000002
8	1796	891	600	1165	305.694	0.0000002
9	1324	874	982	1272	4.836	0.1841764
10	1563	1027	939	923	121.346	0.0000001
11	1351	977	601	1523	239.370	0.0000001
12	1350	897	847	1358	38.920	0.0000001
13	1551	1012	1040	849	153.405	0.0000001
14	1346	1008	584	1514	254.262	0.0000001
15	1201	1007	858	1386	70.008	0.0000001
16	1437	1144	1042	829	190.679	0.0000001
17	1272	967	629	1584	255.494	0.0000002
18	1092	988	968	1404	87.010	0.0000001
19	1233	1127	1253	839	246.561	0.0000002
20	1139	1059	707	1547	232.356	0.0000002
21	1012	1140	1006	1294	149.110	0.0000001
22	1244	1124	1268	816	266.092	0.0000002
23	1194	1012	716	1530	190.022	0.0000001
24	989	1133	1042	1288	159.062	0.0000001
25	1205	1136	1374	737	406.586	0.0000002
26	1129	1054	812	1457	141.482	0.0000001
27	928	1142	1121	1261	210.531	0.0000001

Supplementary Table 16. PWM of nucleotides in *S. enterica* for first 27 nucleotides of other genes in transcription unit

Site	A	C	G	U
1	1.9330	-6.5171	-1.7700	-3.2641
2	-12.1206	-12.1206	-12.1206	2.0983
3	-12.1206	-12.1206	1.8296	-12.1206
4	1.0481	-0.5847	-0.4184	-0.7454
5	0.6094	0.1606	-0.8003	-0.2159
6	0.4779	-0.4486	-0.5589	0.3597
7	0.8309	-0.2752	-0.3145	-0.6314
8	0.7906	-0.3324	-1.0612	0.1643
9	0.3507	-0.3602	-0.3508	0.2911
10	0.5901	-0.1275	-0.4153	-0.1715
11	0.3799	-0.1995	-1.0588	0.5509
12	0.3788	-0.3227	-0.5640	0.3855
13	0.5790	-0.1487	-0.2680	-0.2920
14	0.3745	-0.1544	-1.1002	0.5423
15	0.2101	-0.1558	-0.5454	0.4149
16	0.4689	0.0281	-0.2652	-0.3264
17	0.2929	-0.2143	-0.9932	0.6075
18	0.0729	-0.1833	-0.3715	0.4335
19	0.2480	0.0065	0.0008	-0.3091
20	0.1336	-0.0832	-0.8246	0.5734
21	-0.0369	0.0231	-0.3159	0.3158
22	0.2608	0.0027	0.0179	-0.3492
23	0.2017	-0.1487	-0.8064	0.5575
24	-0.0700	0.0142	-0.2652	0.3091
25	0.2149	0.0180	0.1337	-0.4961
26	0.1209	-0.0901	-0.6249	0.4870
27	-0.1619	0.0256	-0.1598	0.2786

Supplementary Table 17. Background frequency of nucleotides in *M. tuberculosis*

Summary output:

Nuc	A	C	G	T
ObsN	664432	1276445	1332451	684992
Prop	0.1679	0.3225	0.3366	0.1731

GC% = 65.4737
GT% = 50.9353

Supplementary Table 18. Site-specific frequencies of nucleotides in *M. tuberculosis* for first 27 nucleotides

Site	A	C	G	U	X2	p
1	2410	1	1319	176	4999.660	0.0000000
2	0	0	0	3906	15913.793	0.0000000
3	1	2	3902	1	8451.708	0.0000000
4	1604	783	1038	481	1115.691	0.0000004
5	639	1765	868	634	521.029	0.0000003
6	563	1529	1136	678	223.625	0.0000002
7	941	909	1500	556	192.636	0.0000001
8	982	1204	758	962	288.724	0.0000001
9	600	1420	1155	731	125.433	0.0000001
10	972	1133	1277	524	127.842	0.0000001
11	831	1288	771	1016	278.390	0.0000002
12	580	1386	1292	648	133.741	0.0000001
13	886	1016	1433	571	109.258	0.0000001
14	945	1121	763	1077	336.464	0.0000002
15	561	1455	1260	630	184.020	0.0000002
16	853	1059	1473	521	138.350	0.0000001
17	891	1203	798	1014	251.859	0.0000002
18	511	1451	1359	585	244.319	0.0000001
19	781	1123	1526	476	182.244	0.0000002
20	851	1274	742	1039	315.462	0.0000002
21	461	1501	1339	605	300.287	0.0000002
22	727	1181	1542	456	212.167	0.0000001
23	847	1293	785	981	251.197	0.0000002
24	452	1524	1397	533	373.949	0.0000002
25	742	1099	1590	475	219.239	0.0000002
26	946	1147	823	990	234.936	0.0000002
27	463	1505	1420	518	367.811	0.0000002

Supplementary Table 19. PWM of nucleotides in *M. tuberculosis* for first 27 nucleotides of other genes in transcription unit

Site	A	C	G	U
1	1.8774	-9.8959	0.0046	-1.9407
2	-11.9318	-11.9318	-11.9318	2.5300
3	-9.1336	-9.0835	1.5692	-9.1712
4	1.2901	-0.6858	-0.3409	-0.4911
5	-0.0375	0.4865	-0.5988	-0.0928
6	-0.2201	0.2795	-0.2108	0.0040
7	0.5208	-0.4706	0.1901	-0.2821
8	0.5823	-0.0652	-0.7943	0.5086
9	-0.1283	0.1728	-0.1869	0.1125
10	0.5675	-0.1529	-0.0420	-0.3676
11	0.3415	0.0321	-0.7697	0.5874
12	-0.1772	0.1378	-0.0252	-0.0613
13	0.4339	-0.3101	0.1242	-0.2437
14	0.5269	-0.1682	-0.7848	0.6715
15	-0.2252	0.2079	-0.0614	-0.1019
16	0.3792	-0.2503	0.1639	-0.3759
17	0.4420	-0.0664	-0.7201	0.5846
18	-0.3599	0.2039	0.0477	-0.2088
19	0.2520	-0.1657	0.2149	-0.5062
20	0.3758	0.0163	-0.8250	0.6197
21	-0.5084	0.2528	0.0264	-0.1603
22	0.1486	-0.0930	0.2299	-0.5681
23	0.3690	0.0376	-0.7438	0.5368
24	-0.5368	0.2747	0.0875	-0.3431
25	0.1781	-0.1968	0.2742	-0.5092
26	0.5284	-0.1352	-0.6756	0.5500
27	-0.5021	0.2566	0.1111	-0.3842

Supplementary Table 20. Genes separated by position in transcription unit

List of first translated genes in a transcription unit and the other genes in the same transcription units alongside the first 27 nucleotides of their sequences. A total of 6 transcription units were found from RegulonDB.

First gene	Sequence
rpmF	ATGGCCGTACAACAGAATAAACCAACC
rpmI	ATGCCAAAAATTAAGACCGTACGCGGT
rpsP	ATGGTAACTATTCGTTTAGCACGTCAC
rplU	ATGTACGCGGTTTTCCAAAGTGGTGGT
rpmB	ATGTCCCGAGTCTGCCAAGTTACTGGC
rplJ	ATGGCTTTAAATCTTCAAGACAAACAA
Other genes	
plsX	TTGACACGTCTAACCCTGGCGTTAGAT
fabD	ATGACGCAATTTGCATTTGTGTTCCCT
fabG	ATGAATTTTGAAGGAAAAATCGCACTG
rplT	ATGGCTCGCGTAAAACGTGGTGTTATT
rplS	ATGAGCAACATTATTAAGCAACTTGAA
trmD	ATGTGGATTGGCATAATTAGCCTGTTT
rimM	ATGAGCAAACAACTCACCGCGCAAGCA
obgE	ATGAAGTTTGTTGATGAAGCATCGATT
yhbE	ATGAAGCAGCAGGCAGGCATTGGCATT
rpmA	ATGGCACATAAAAAGGCTGGCGGCTCC
mutM	ATGCCTGAATTACCCGAAGTTGAAACC
rpmG	ATGGCTAAAGGTATTCGTGAGAAAATC
rplL	ATGTCTATCACTAAAGATCAAATCATT
rpoB	ATGGTTTACTCCTATACCGAGAAAAAA
rpoC	GTGAAAGATTTATTAAAGTTTCTGAAA