12-1-2007

The tri-frame model

Elsje Pienaar
University of Nebraska - Lincoln, epienaar2@unl.edu

Hendrik J. Viljoen
University of Nebraska - Lincoln, hviljoen1@unl.edu

www.elsevier.com/locate/yjtbi

The tri-frame model

Elsje Pienaar, Hendrik J. Viljoen

PII: S0022-5193(07)00620-0
DOI: doi:10.1016/j.jtbi.2007.12.003
Reference: YJTBI 4964

To appear in: *Journal of Theoretical Biology*

Received date: 17 July 2007
Revised date: 7 December 2007
Accepted date: 7 December 2007

Cite this article as: Elsje Pienaar and Hendrik J. Viljoen, The tri-frame model, *Journal of Theoretical Biology* (2007), doi:10.1016/j.jtbi.2007.12.003

THE TRI-FRAME MODEL

By

Elsje Pienaar and Hendrik J.Viljoen*

Department of Chemical and Biomolecular Engineering

University of Nebraska, Lincoln, NE 68588-0643

***Corresponding Author. Mailing address: 211 Othmer Hall, University of Nebraska 68588-0643, Phone: 1(402) 472-9318, Fax: 1(402) 472-6989, E-mail: hviljoen1@unl.edu**

Abstract

The tri-frame model gives mathematical expression to the transcription and translation processes, and considers all three reading frames. RNA polymerases transcribe DNA in single nucleotide increments, but ribosomes translate mRNA in pairings of three (triplets or codons). The set of triplets in the mRNA, starting with the initiation codon (usually AUG) defines the open reading frame (ORF). Since ribosomes do not always translocate three nucleotide positions, two additional reading frames are accessible. The -1RF and the +1RF are triplet pairings of the mRNA which are accessed by shifting one nucleotide position in the 5' and 3' directions respectively. Transcription is modeled as a linear operator that maps the initial codons in all three frames into other codon sets to account for possible transcriptional errors. Translational errors (missense errors) originate from misacylation of tRNA's and misreading of aa-tRNA's by the ribosome. Translation is modeled as a linear mapping from codons into aa-tRNA species, which includes misreading errors. A final transformation from aa-tRNA species into amino acids provides the probability distributions of possible amino acids into which the codons in all three frames could be translated. An important element of the tri-frame model is the ribosomal occupancy probability. It is a vector in R^3 that gives the probability to find the ribosome in the ORF, -1RF or +1RF at each codon position. The sequence of vectors, from the first to the final codon position, gives a history of ribosome frameshifting. The model is powerful: it provides exact expressions for: (1) yield of

error-free protein, (2) fraction of prematurely terminated polypeptides, (3) number of transcription errors, (4) number of translation errors and (5) mutations due to frameshifting. The theory is demonstrated for the three genes *rpsU*, *dnaG* and *rpoD* of *E. coli* which lie on the same operon, as well as for the *prfB* gene.

Key words: Mathematical model; Transcription; Translation; Frameshifting; Error prediction.

Introduction

Transcription and translation can be illustrated by the sequential steps: DNA → mRNA→proteins: DNA polymerases catalyze the copying of DNA, RNA polymerases are responsible for the transcription of DNA into mRNA and ribosomes perform the complex functions to translate the mRNA sequence and synthesize new proteins. The DNA polymerases and RNA polymerases process their templates one nucleotide at a time, but the ribosomes translate the mRNA in multiples of three nucleotides, usually referred to as codons or triplets. The processing of three nucleotides at a time requires three reading frames to be considered: the open reading frame (ORF), the +1RF and the -1RF; respectively defined as the set of triplets that coincide with the initiation codon (usually $[AUG]$), the set that is shifted one nucleotide position in the 3' direction with respect to ORF and the set that is shifted one nucleotide position in the 5' direction. The two main objectives of this study are to give mathematical expression to the transcription and translation processes, with specific emphasis on the loss of fidelity, and to consider all three reading frames in the analysis.

The standard genetic code of Nirenberg *et al.* (1966) assigns 64 RNA triplet code words for 20 amino acids and a translational stop. Since the 1960's most researchers have focused primarily on how an amino acid sequence is decoded from mRNA in one reading frame. The successful synthesis of a protein requires that the ribosome must accurately translate messenger RNA in the correct frame. Most genes code only for single proteins. But ribosomes may still slip by one base in either the 3' (+1) or 5' (-1) direction and translate mRNAs out-of-frame. In most cases these frameshifting events lead to out-of-frame termination and the polypeptide chains serve no other purpose but to be tagged for destruction and later destroyed. However, overlapping, same-sense genes code for proteins in two different frames and occasional frameshifting at specific sites are intentional. An interesting example where frameshifting is used to access genetic information in another frame is the *dnaX* gene of *E.coli*. The *dnaX* gene codes for the

τ subunit and the γ subunit of the DNA polymerase of *E.coli*. Both proteins are encoded in the 0RF, but in the case of the γ subunit, a -1 frameshift occurs at the 431st codon to cause early termination. The *prfB* gene of *E.coli* codes for release factor 2 that facilitates translational termination at the UGA and UAA stop codons. A UGA stop codon at the 26th codon position in the ORF would normally cause early termination, but at low concentration of release factor 2, the ribosome shifts to the +1 frame, which contains the remainder of the encoded sequence. If the release factor concentration increases, early termination at the 26th codon position becomes more likely.

There is strong evidence that ribosome pause times govern frame shift frequencies and the availability of cognate tRNA influences this process (Sipley and Goldman, 1993). Another factor that affects frameshifting is secondary structures in the mRNA, such as knots and stem loops (Farabaugh, 1997; Tsuchihashi, 1991). On average, 27% to 31% of *E.coli* β-galactosidase mRNA molecules terminate prematurely during translation (Lindsley *et al*., 2005; Manley, 1978), although reading frame (RF) error rates for completely translated mRNA are much lower. Kurland (1979) and Marquez *et al.* (2004) have measured the reading frame error rate in *E.coli* as approximately 3×10^{-5} per codon.

It is interesting to note how out-of-frame translation is terminated. Translation in the +1 reading frame is terminated by stop codons that form if the in frame RNA code words for Leucine (CUG, CUA, UUG, UUA), Valine (GUG, GUA), Isoleucine (AUA) or Methionine (AUG) are followed by an A or G. Thus the triplet amino acid code words L, V, I and M overlap translational stop code words UGA, UAA and UAG. The frequencies of amino acids in proteins generally occur in the order L>A>G>S>V>E>K>T>P>D>R>I>N>Q>F>Y>H>M>C>W (Cserzo and Simon, 1989; cf. order of amino acids listed in Table 1). Leucine is the most common amino acid in all protein data bases, and four of the six Leucine codons can form translational stops in the +1 reading frame when followed by an A or G. For example, the most frequent RNA code word for Leucine in most

organisms is CUG (Andersson and Kurland, 1990; Ikemura, 1985; Sharp *et al*., 1988). If CUG is followed by a 3' A, then a translational stop CUGA results in the +1 reading frame. The RNA code words for A, G, S, V, E, K and T, which are the next most frequent amino acids in proteins, all begin with A or G. Therefore, amino acid code words with translational stops embedded in their 2nd and 3rd codon positions (NUG and NUA) are most likely followed by a purine due to the frequent occurrence of A or G in the adjacent amino acid: NUGA, NUAA or NUAG (see Table 1).

RNA code words that begin with AA, AG and GA can overlap translational stops in the -1 reading frame if they are preceded by a 5' U. These code words encode Lysine (AAG, AAA), Arginine (AGG, AGA), Glutamic acid (GAG, GAA), Asparagine (AAU, AAC), Aspartic acid (GAU, GAC) and Serine (AGU, AGC). Therefore these amino acids are protected from mistranslation in the -1 reading frame. For example, if Lysine AAG or AAA codons are preceded by a 5' U, then the RNA sequences UAAG or UAAA result; and translational stops are thus encoded in the –1 reading frame (Table 1). In general, when amino acid code words with translational stops embedded in their 1st and 2nd positions (GAN, AGN and AAN) are preceded by a 5' U, translational stops are encoded in the -1 reading frame: UGAN, UGAN or UAAN. In single letter code, the amino acids L, A, G, V, T, P, D, R, I, N, F, Y, H and C each have one triplet RNA code word that ends with a 3rd position U, and two of the six Serine codons end with a 3rd position U. Thus the -1 reading frame stops are programmed to occur relatively frequently.

At least six research groups have previously recognized that there must be some kind of error control mechanism in order to avoid out-of-frame translation (Antezana and Kreitman, 1999; Archetti, 2004; Hansen *et al*., 2003; Marquez *et al*., 2005; Seligmann and Pollock, 2004; Stahl *et al*., 2002). Furthermore, Konopka (1985) has shown that the degeneracy of the genetic code provides some error protection during transcription. The difference between the information entropy at the input (mRNA) and the output (amino acid sequence) is a measure of the degree

of error protection. Antezana and Kreitman (1999) considered the role out-of-frame codons could play and stated, “The statistically significant congruency of in-frame and off-frame trinucleotide preferences suggests that the same kind of reading frame independent force(s) may also influence synonymous codon choices.”

Hansen *et al.* (2003) have described an elegant mechanism by which translational error control is achieved on the ribosome: “…the translational frame is controlled mainly by the stability of codon-anticodon interactions at the A-site.” Harger *et al.* (2002) have proposed a kinetic model termed the “integrated model” of programmed ribosomal frameshifting. In this model, the kinetics of protein translation are simplified into four stages: (1) selection and insertion of aminoacyl-tRNA into the ribosomal A-site, (2) accommodation of the 3’ end of the aminoacyl-tRNA into the P-site, (3) peptidyl transfer, and (4) translocation. The aminoacyl-tRNA occupancy states of the ribosome are different in +1 reading frames, as compared to –1 reading frames. Only the first accommodation step involves the ribosomal A-site. Therefore, according to the model, the shift to the +1 reading frame occurs when the A-site is empty, whereas the shift to the -1 reading frame occurs when both the A- and P-sites are occupied.

A mathematical model is presented of transcription and translation. All three reading frames are considered and the ribosome may access other frames – thus the concept of ribosomal occupation distribution is introduced. The model shows that errors occur during translation and during transcription, but the degeneracy of the genetic code provides some protection against these errors. The model demonstrates that variations in the translation rates of different codons and termination of non-programmed frameshifting events are mechanisms of posttranscriptional control.

Elements of the mathematical model

The general approach

The tri-frame theory is a mathematical expression of the process illustrated by:

DNA →RNA →Protein. (1a)

The tri-frame theory links the three possible reading frames in mRNA by the mechanism of ribosome frameshifting. The theory offers new insight into the process of encoding that is used by the DNA to ensure that a protein of specific amino acid composition is synthesized. The theory further describes post-transcriptional modulation of synthesis levels and the control of accuracy of the product.

The DNA and mRNA are directionally processed, from the 3' to 5' end and from the 5' end towards the 3' end respectively. Since the RNA polymerase transcribes the DNA one nucleotide at a time and therefore translocates in single nucleotide steps, transcription is frame insensitive. The ribosome translates and translocates three nucleotides at a time, thus three frames are identified with the process. The open reading frame (ORF) is defined as the set of triplets (or codons) which start with the initiation codon, usually AUG. It is the intended frame the ribosome ought to process. The -1RF defines the set of triplets by shifting one nucleotide position in the 5' direction, the +1RF is obtained by a single shift in the 3' direction (+1RF). For the development of the model, it is necessary to introduce the codon description already at the transcription stage. Therefore the DNA sequence is grouped into the three frames. Full details of the mathematical model only follow hereafter, but it is helpful to introduce some notation. Starting with the DNA, the sequence is considered as three parallel sets of sequential codons, namely the set in ORF together with the alternative sets in the $\pm$1RF's. The codons in all three frames at the *i*th position are uniquely identified by the matrix C^i. The transcription process is mathematically described by the matrix T and the transcribed codons in all three frames are designated D^i. The matrix M describes the translation process and S^i is the matrix of translated codons. Multiplication by the (Nirenberg) transformation matrix N_S maps S^i into the matrix S^i_{AA} that contains the amino

acid composition at the *ith* position. The mathematical operations and the equivalent biochemical steps are shown in expression (1b).

ith codons in DNA $\rightarrow$ *ith codons in mRNA* $\rightarrow$ *amino acid(s) at ith position in polypeptide chain*

$$C^i \quad \times T \quad \rightarrow \quad D^i \quad \times M \quad \rightarrow \quad [\, S^i \times N_S \,] \rightarrow S^i_{AA} \qquad (1b)$$

Translation occurs only in one frame, but the ribosome may switch between frames (Weiss *et al.* (1990)). In parallel to (1b) is the process of ribosome frameshifting and it plays the very important role to connect the information encoded in the three reading frames. Frameshifting is not a deterministic process. Since pausing of the ribosome at codons that translate slowly, increases the probability of frameshifting, the process is of stochastic nature. We introduce the vector P^i that consists of three probabilities, to describe the likelihood that the ribosome may be in a specific reading frame. It is referred to as the ribosome occupancy distribution. Let V be a matrix that contains the frameshifting probabilities of all the codons. Then $[D^i \times V]^T P^{i-1} = P^i$ is the mathematical equation that describes what the ribosome occupancy distribution will be if frameshifting occurs during translation of the *i*th codons (three frames).

Transcription

Consider a segment of a DNA molecule that codes for a protein and let its open reading frame consist of 3N base pairs. Pair the bases of the open reading frame into triplets and index the codons: c_i, $i = 1,..N$. There are 64 different codons, including the three stops. Assign number values to the nucleotides as follows: $T = 1; C = 2; G = 3; A = 4$. Thence a generic triplet IJK at the *ith* codon position is identified by an index between 1 and 64; define the index of a codon at the *ith* position (there are three codons at the *ith* position) as $c_i = 4^2(I-1) + 4(J-1) + K$. The identity of the *ith* codon is expressed in terms of a vector as follows:

$\overline{\sigma}^{i} = \{\sigma^{i}(1), \sigma^{i}(2)..,\sigma^{i}(j),..\sigma^{i}(64)\}$, where $\sigma^{i}(j) = 0$ if $j \neq c_i$, and $\sigma^{i}(c_i) = 1$. In the same manner that index c_i labels the *ith* codon in 0RF, codons in the $\pm$ 1RF are labeled by c_i^{+} and c_i^{-} respectively. The out-of-frame codons are represented by the vectors $\overline{\sigma}^{-i}$ and $\overline{\sigma}^{+i}$. We combine the vectors of all three reading frames to form the 3X64 matrix;

$$C^{i} = \begin{bmatrix} \overline{\sigma}^{-i} \\ \overline{\sigma}^{i} \\ \overline{\sigma}^{+i} \end{bmatrix}. \qquad (2)$$

Each row of C^{i} is a vector that must be interpreted as a probability distribution over 64 codons. Therefore the implication of $\sigma^{i}(c_i) = 1$ in each row of eq.(2) is that the initial data, in other words the DNA information, is presented with hundred percent certainty.

The index c_i is uniquely mapped to an amino acid a_i (the inverse mapping is not unique). We number the amino acids, using their one letter symbols, in the order: L=1; A=2; G=3; S=4; V=5; E=6; K=7; T=8; P=9; D=10; R=11; I=12; N=13; Q=14; F=15; Y=16; H=17; M=18; C=19; W=20; X=21. For example, if the third codon is $[AGT]$, then $c_3 = 4^2(4-1) + 4(3-1) + 1 = 57$ and $a_3 = 4$ (Serine). The vector $\overline{\sigma}^{3}$ is: $\overline{\sigma}^{3} = \{000...1^{(57)}0000000\}$, where the superscript (57) denotes the column position.

Transcription is not error-free, there is a small probability that a nucleotide is mistranscribed. The 64X64 matrix $T = \{t(i,j)\},\ i,j = 1..64$ consists of the probabilities to mistranscribe. Thus $t(i,j)$ is the probability that a codon with index *i* is transcribed as a codon with index *j.* Konopka (1985) assumed that only one mistranscription can occur for any triplet, consequently there are nine incorrect possibilities for each triplet. There are 27 possibilities for two mistranscriptions per triplet and 27 possibilities that all three nucleotides of a triplet are mistranscribed. Let β denote

the probability to mistranscribe a nucleotide into another one. (If information on nucleotide-specific mistranscription is known, it is straightforward to include that information.) Each row of T has nine linear elements $t(i,j)=\beta$, $i \neq j$, twenty seven quadratic elements $t(i,j)=\beta^2$, $i \neq j$ and twenty seven cubic elements $t(i,j)=\beta^3$, $i \neq j$ for a total of sixty three different mistranscriptions. In theory all codons are accessible by transcription, with varying probabilities. The diagonal element $t(i,i)=1-9\beta-27\beta^2-27\beta^3$ is the probability to transcribe correctly. The sum of each row of *T* is one. In mathematical terms, transcription is described by multiplying C^i with *T*;

$$[C^i \text{ X } T] = D^i = \begin{bmatrix} \bar{d}^{-i} \\ \bar{d}^{i} \\ \bar{d}^{+i} \end{bmatrix}. \qquad (3)$$

D^i is a 3X64 matrix and its top, middle and bottom rows correspond to the probability distributions of the *ith* codon in the -1RF, 0RF and +1RF respectively. The sum of elements in each row of D^i is one, since it presents all possible transcription outcomes. Of significance is the fact that the original codon information is no longer present with certainty. For example, if $c_i = 3$ (i.e. the *ith* codon is $[TTG]$), $c_i^- = 49 = [ATT]$ and $c_i^+ = 11 = [TGG]$, then

$$D^i = \begin{bmatrix} \beta^{(1)}..\beta^{(17)}..\beta^{(33)}..(1-9\beta-27\beta^2-27\beta^3)^{(49)}\beta^{(50)}\beta^{(51)}\beta^{(52)}\beta^{(53)}..\beta^{(57)}..\beta^{(61)}.. \\ \beta^{(1)}\beta^{(2)}(1-9\beta-27\beta^2-27\beta^3)^{(3)}\beta^{(4)}...\beta^{(7)}...\beta^{(11)}...\beta^{(15)}..\beta^{(19)}..\beta^{(35)}..\beta^{(51)}.. \\ \beta^{(3)}..\beta^{(7)}..\beta^{(9)}..\beta^{(10)}(1-9\beta-27\beta^2-27\beta^3)^{(11)}\beta^{(12)}..\beta^{(15)}..\beta^{(27)}..\beta^{(43)}..\beta^{(59)}. \end{bmatrix}. \qquad (4)$$

Notes: (a) The superscripts in eq.(4) denote column positions. (b) Only first order errors are indicated in eq. (4), except the index positions that include second and third order errors.

The probability that the *ith* codon $[TTG]$ is transcribed to $[UUG]$ is $D^i(2,3) = 1 - 9\beta - 27\beta^2 - 27\beta^3$, but the probability that it is mistranscribed to $[UUC]$ is $D^i(2,2) = \beta$.

Translation

In a review by Parker (1989) two sources of mistranslation are discussed. The first source is misacetylation of tRNA's. The average frequency with which aminoacyl-tRNA synthetases charge tRNA incorrectly varies between 4×10^{-4} and 5×10^{-5}. Closely related amino acids are substituted for the correct one. The second source of mistranslation is misreading, which implies incorrect binding of an aa-tRNA to the A site of the ribosome.

Kramer and Farabaugh (2007) experimentally determined the frequency of misreading errors for each one of the fourteen near-cognates for the two codons of lycine, $[AAG]$ and $[AAA]$. The frequencies varied from 3.1×10^{-4} to 36×10^{-4}; the two codons most frequently misread by $tRNA^{Lys}_{UUU}$ are $[AGA]$ and $[AGG]$. For example, the codon $[ACG]$ that codes for tyrosine, is misread as a lysine with frequency 3.1×10^{-4}, but the codon $[AGG]$, that codes for arginine, is misread for lysine ten times more, 31×10^{-4}. Kramer and Farabaugh noted that the rare mutants $[AGG]$ and $[AGA]$ are misread as lysine ten times more than the other near –cognates, an observation that correlates strongly with the availability of their tRNA. These experimental data are valuable, but they are not complete. Due to the paucity in experimental data, it is necessary to obtain theoretical estimates of the misreading frequencies.

Estimation of misreading frequencies

Near-cognate aa-tRNAs are defined to have a single mismatch in the codon-anticodon loop in either the 2nd or 3rd position. Since some cognate tRNAs have a mismatch in the 3rd position,

these tRNAs are excluded from the set of near-cognates. The binding of aa-tRNA to the A site is the first step in the kinetics of peptide synthesis by the ribosome and there are further editing and proofreading steps which determine ultimately if an amino acid is transferred to the nascent peptide or not. In a recent study, Fluitt *et al.* (2007) used the kinetic model and experimentally determined rate constants of Gromadski and Rodnina (2004) to derive an expression for the average insertion time of an amino acid in the peptide chain from a cognate aa-tRNA. The average time to translate a codon at the *ith* position (in *ms*) is:

$$\tau = 9.06 + 1.445 \times \left[10.48C(c_i) + 0.5R(c_i)\right] \quad (5)$$

The insertion time is delayed by competition from near-cognates and non-cognates. The competition measures (C and R) depend on the codon index c; their definitions are as follows:

$$C(c) = \frac{\sum_{k \in Near-C} (t_k)^{-1}}{\sum_{m \in Cog} (t_m)^{-1}}, \; c = 1 \ldots 64, \quad (6a)$$

$$R(c) = \frac{\sum_{k \in Non-C} (t_k)^{-1}}{\sum_{m \in Cog} (t_m)^{-1}}, \; c = 1 \ldots 64, \quad (6b)$$

Near-C and *Non-C* are the sets of near-cognate and non-cognate aa-tRNAs respectively and *Cog* is the set of cognate aa-tRNAs for the codon with index c.

In order to apply eq. (6a,b), one must calculate the arrival times t_k of the different tRNA's (see Fluitt *et al*. (2007) for details). The arrival times are the average times it takes aa-tRNA complexes to diffuse towards the A site of the ribosome. The values are based on the amount of tRNA available in a cell (we used values at the logarithmic phase at a growth rate of 0.4 doublings per hour) and the average number of ribosomes which are actively translating. The inverse of the arrival times are the arrival rates. The tRNA species and release factors are listed in Table 2 together with their average number/cell (Dong *et al.* (1996)) and their arrival times.

The probability to insert an incorrect amino acid into the nascent peptide chain is directly proportional to the number of binding attempts by near-cognates. Based on this approach the values in Table 3 have been obtained. The 64 codons (including the three stops) are translated by 46 tRNA's and two terminating factors. Table 3 lists the codons, the misread amino acid and the frequency of that occurrence. The competition measures as defined by eqns (6a,b) are also included in Table 3.

Eqns (5, 6a and 6b) are results of a comprehensive mathematical model of ribosomal kinetics and translational fidelity, described only briefly here. Interested readers are referred to Fluitt *et al*. (2007) for a more detailed description of the underlying model.

Transformation matrix M

The 64X48 transformation matrix M maps the transcribed matrix D^i into the translated matrix S^i. Each row of M corresponds to a specific codon and $m(i,j)$ is the probability that a codon with index *i* is translated by the *jth* aa-tRNA (note that *j=47,48* correspond to release factors). The labels $j = 1...48$ that define the aa-tRNA species (i.e. columns of M) are listed in Table 2. Since every codon is eventually translated, the sum of probabilities in any row of M should be one.

To illustrate the point, consider the codon $[ACG]$ which codes for threonine and its index is $c = 55$. There are two cognate tRNA's, namely $tRNA^{Thr}_{UGU}$ and $tRNA^{Thr}_{CGU}$. The near-cognate tRNA's that only mismatch in the 3rd position are *Thr1, Thr3,* both codes for threonine. The near-cognate tRNA's that mismatch in the 2nd position are *Arg5, Ile2, Metf1, Metf2, Metm*. Thus the non-zero components of the 55th row are: *m(55,40), m(55,38), m(55,37), m(55,39), m(55,6), m(55,18), m(55,25), m(55,26), m(55,27)*. The respective amino acids are: T, T, T, T, R, I, M, M,

M. If translation is error-free, then only the cognate tRNA's have non-zero values, i.e. *m(55,40)* and *m(55,38)*. Furthermore, their sum should be one, *m(55,40)* + *m(55,38)* = 1. If misreading is considered, then it follows from Table 3 that *m(55,6)* $= 2\times10^{-4}$, *m(55,18)* $= 8\times10^{-4}$, *m(55,25)* $= 5\times10^{-4}$, *m(55,26)* $= 2\times10^{-4}$ and *m(55,27)* $= 3\times10^{-4}$ (Table 3 lists the sum of all methionine species). The sum of the remaining values of row 55 is $1-20\times10^{-4}=0.9980$. The mathematical model of Fluitt *et al.* (2007) provides the values of the cognate and near-cognates which translate threonine, specifically *m(55,37)*= 4×10^{-5}, *m(55,38)*= 0.3685, *m(55,39)*= 5×10^{-4} and *m(55,40)*= 0.6289. The probability that one of the two near-cognates (*Thr1* or *Thr3*) translates threonine is small, but the cognates $tRNA^{Thr}_{UGU}$ and $tRNA^{Thr}_{CGU}$ have probabilities 0.6289 and 0.3685 respectively.

The translated matrix S^i

Multiply D^i with M to obtain S^i. The translation process is mathematically expressed by

$$D^i \text{XM} = S^i = \begin{bmatrix} \overline{s}^{-i} \\ \overline{s}^{i} \\ \overline{s}^{+i} \end{bmatrix} \qquad (7)$$

S^i is a 3X48 matrix. Rows 1,2 and 3 of S^i give the tRNA distributions of the -1, 0 and +1 reading frames respectively. To obtain the amino acid distribution, S^i is multiplied with a matrix that relates tRNA's to amino acids (the first three columns of Table 2 provides the information for this transformation).

Protein Composition

At any stage of the process, following either transcription or translation, the D^i or S^i matrix can be multiplied with transformation matrix *N* or N_S respectively to obtain the amino acid distribution at

the *ith* position. The product D^i x *N* (in this case *N* is a 64x21 matrix) is interpreted as the amino acid probability distribution *if no errors occur during translation*.

$$D^i_{aa} = D^i \times N \tag{8}$$

The product S^i x N_S (in this case *N* is a 48x21matrix) is the amino acid probability distribution after the translation step:

$$S^i_{aa} = S^i \times N_S \tag{9}$$

Note that the elements $D^i_{aa}(2, a^i)$ and $S^i_{aa}(2, a^i)$ mark the probabilities of actually adding the amino acid a^i that the codon with index c^i has coded for, into the nascent polypeptide chain. Transcription and translation errors spread the distribution, whilst degeneracy tends to focus the distribution.

The Ribosomal Occupancy of the Three Reading Frames

A complication that has not been addressed until now is frameshifting. The ability of the ribosome to frameshift and translate in any one of three frames is the reason why the probability distributions are presented in all three reading frames, hence three rows in matrices C^i, D^i and S^i. The translational process is interrupted if the ribosome detaches from the mRNA, or if a stop codon is encountered. Stop codons are usually encountered at the end of translational process. Occasionally, ribosomes slip by one base in either the 3' (+1) or 5' (-1) direction and translate mRNAs out-of-frame. Following this event, there is a high probability that a stop codon is encountered and the translational process terminates prematurely.

We assign the probability ψ_c for the ribosome to frameshift at the codon with index *c*. If certain putative sequences promote frameshifting either by forming secondary structures that hinder ribosomal translation, or slippery sites that affect frame integrity, and these effects can be quantified in terms of probabilities, the value ψ_c values can be updated accordingly. The problem with sequence dependent frameshifting is lack of quantitative data. To keep the model general, a distinction is made between frameshifting in the 5' direction or the 3' direction, specifically we denote the probability to shift towards the 5' end as γ^- and the probability to shift towards the 3' end as γ^+. Thus the probability to frameshift at a codon of index *c* is $\psi_c = \gamma_c^- + \gamma_c^+$. The probability that the ribosome remains in the current frame is α_c, the probability that the ribosome detaches from the mRNA prematurely is μ_c and the sum of these outcomes is one;

$$\psi_c + \alpha_c + \mu_c = 1.$$

Let V be a 64X3 matrix that contains the frameshifting probabilities for all 64 codons. The kth row of V is defined as $\left[\gamma_k^- \alpha_k \gamma_k^+\right]$ and it consists of the probabilities of the ribosome to frameshift at a codon with index k in the 3' direction, to remain in the current frame or to frameshift to the 5' direction. Therefore the vectors $\overline{\gamma}^-$, $\overline{\gamma}^+$ and $\overline{\alpha}$ consist of the probabilities to frameshift in the 5' and 3' directions or remain in frame for all 64 codon indices. The vectors form the columns of V as follows:

$$V = \left[\overline{\gamma}^- \quad \overline{\alpha} \quad \overline{\gamma}^+\right] \qquad (10)$$

Note: The α_k values for rows 12,15 and 16 of V are zero since they correspond to the stop codons. If termination factor is present in low molar fractions, the ribosome may frameshift at these codons. Practically, the occurrence of stop codons is limited to the +1RF and -1RF (with rare exceptions, stop codons generally only appear at the end of the 0RF).

The matrix D^i contains the probabilities distributions of the *ith* codons in all three frames. The product of any row of D^i with the 2nd column of V is the probability that the ribosome remains in the frame that corresponds to that row. Likewise the products of any row of D^i with the 1st or 3rd columns of V are the probabilities to shift towards the 5' or 3' directions with respect to the corresponding frame. The results are presented in the 3X3 matrix R^i.

$$R^i = D^i \times V = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} = \begin{bmatrix} \bar{d}^{-i} \cdot \bar{\alpha} & \bar{d}^{-i} \cdot \bar{\gamma}^{+} & \bar{d}^{-i} \cdot \bar{\gamma}^{-} \\ \bar{d}^{i} \cdot \bar{\gamma}^{-} & \bar{d}^{i} \cdot \bar{\alpha} & \bar{d}^{i} \cdot \bar{\gamma}^{+} \\ \bar{d}^{+i} \cdot \bar{\gamma}^{+} & \bar{d}^{+i} \cdot \bar{\gamma}^{-} & \bar{d}^{+i} \cdot \bar{\alpha} \end{bmatrix}. \quad (11)$$

The matrix R^i does not only contain the probabilities which determine the ribosome occupancy behavior, but also links the process of encoding to the conditions in the cell. If aa-tRNA pool compositions change, the pause times and hence frameshifting probabilities are affected.

Next we calculate the occupancy probabilities of the ribosome. The values p^{-i}, p^i and p^{+i} are the probabilities that translation at the *i+1th* codon position occurs in the -1RF, 0RF or +1RF:

$$P^i = \begin{bmatrix} p^{-i} \\ p^i \\ p^{+i} \end{bmatrix}. \quad (12)$$

It is assumed that the ribosome is properly aligned with the zero reading frame when protein synthesis is initiated, therefore $P^0 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$. (13)

The probability P^i is calculated as follows;

$$\left[R^i\right]^T P^{i-1} = P^i \quad (14)$$

The set $\{P^i, i = 0,1,2..N\}$ describes the probability that the ribosome is in a specific frame for a codon at the *ith* position. An alternative interpretation is to consider a large number of ribosomes, processing similar mRNAs. The set $\{P^i, i = 0,1,2..N\}$ presents the (normalized) average number of ribosomes that occupy each frame at the *ith* codon position.

Useful expressions of the tri-frame model

The probability that no frameshifting has occurred at any one of the N codons in the mRNA is given by

$$\eta = \prod_{i=1}^{N} \bar{d}_i \cdot \bar{\alpha} \,. \tag{15}$$

Return to eq.(9) for a moment, S_{aa}^i represents the amino acid distributions in all three reading frames after translation at the *ith* codon position. The product

$$A^i = \left[S_{aa}^i\right]^T \times P^i \tag{16}$$

is a 21-vector that represents the amino acid probability distribution at the *i*th codon position. Therefore the overall protein probability distribution is given by the set $\{A^i, 1 \le i \le N\}$.

The total protein yield without any frameshift or mistranslation errors is;

$$\upsilon = \prod D_{aa}^i(2, a^i)\left[\bar{d}_i \cdot \bar{\alpha}\right] \tag{17}$$

Along the same line of reasoning, the total protein yield without errors of any kind is;

$$\xi = \prod S_{aa}^i(2, a^i)\left[\bar{d}_i \cdot \bar{\alpha}\right] \tag{18}$$

The tri-frame coding theory provides several important results which are summarized in Table 4.

Application of the Tri-frame Model

The theory is applied to four genes of *E.* coli: *prfB*, *rpsU, rpoD,* and *dnaG*. First consider the latter three genes. The genes *rpsU* and *rpoD* flank the *dnaG* gene on the 5' and 3' sides and the three genes all belong to a single macromolecular synthesis operon. Konigsberg and Godson (1983) did DNA sequencing of the genes and found that the *dnaG* primase gene uses an unusually large number of rare codons. Typically the codons *AUA, UCG, CCC, ACG, CAA, AAU* and *AGG* appear only 4% in the zero reading frame and 11% and 10% in the non-reading frames. In the case of *dnaG*, these rare codons appear 11% in the zero reading frame and 12% in the non-reading frames. Konigsberg and Godson suggested that translational modulation using isoaccepting tRNA availability may be part of the mechanism to keep *dnaG* gene expression low. The argument is extended to the repressor genes *lacI, araC* and *rpsR* which also use rare codons, and a general mechanism is proposed that the cell uses rare codons to modulate protein product levels that cannot be tolerated in the cell in excess amounts. The DNA sequences of the open reading frames of *rpsU, dnaG* and *rpoU* have been obtained from Genbank and are provided as Supplementary Material. The *rpsU* gene codes for 72 amino acids, the *dnaG* gene codes for 582 amino acids and the *rpoD* gene codes for 614 amino acids.

Experimental data for frameshift probabilities are not available, but the argument based on ribosome pause time provides a method to estimate the values for a phenomenological evaluation of the model. The time that elapses between filled states of the ribosomal A site is referred to as the pause time. The longer the pause time, the more likely the ribosome is to shift frames. We propose that the pause times, and hence the frameshift probabilities, are proportional to the number of non-cognate binding attempts. The competition measures from non-cognates are normalized and scaled by factor *k* to obtain the frameshift probabilities:

$$\psi_c = R(c)/(k\sum R_c). \quad (19)$$

If γ_c^- and γ_c^+ are the probability to shift either towards the 5' end or the 3' end, then $\psi_c = \gamma_c^- + \gamma_c^+$. The probability to stay in-frame is:

$$\alpha_c = 1 - R(c)/(k\sum R_c) \quad (20)$$

In the application that follows, we assign equal probabilities to γ_c^- and γ_c^+:

$$\gamma_c^+ = \gamma_c^- = 0.5\psi_c. \quad (21)$$

If information about codon-specific frameshift bias becomes available, then γ_c^- and γ_c^+ can be updated accordingly.

We have used $k = 500$ in this study, because this value gives an average frameshift probability per codon of 3 x 10^{-5}, which is consistent with the reading frame error rate in *E.coli* which has been measured by Kurland (1979) and Marquez *et al.* (2004).

Results for the rpsU gene

The *rpsU* gene is relatively small, it has 72 codons. The matrices C^i, $i = 1...72$ are set up according to eq.(2). The transcription error rate of $\beta = 3\times10^{-4}$ has been used (cf. Konopka (1985)). The error frequencies which are used in the transformation matrix *M* are given in Table 3. The frameshift probabilities have been calculated as described in eqns. (19-21).

In Figure 1 the ribosome occupancy distribution is shown as a function of the codon position in the mRNA of the *rpsU* gene. The ribosome remains primarily in the ORF and the probability that it is still in frame at the end is $P^{N-1}(2) = 0.992$. The out-of-frame occupancies show sudden reductions to zero at positions where out-of-frame stop codons have caused the termination of translation. The sum of all three occupancy probabilities is not necessarily one, due to out-of-frame terminations.

The probability that the ribosome never frameshifts is given by eq. (15). For the *rpsU* gene $\eta = 0.992$. That means that in 99.2% of all cases the full-length protein is synthesized without frameshifting. The fraction (of all translational attempts) that terminates out-of-frame is denoted by $\Gamma = 1 - P^{N-1}(2)$. In this case the early terminations account for $\Gamma = 1 - 0.992 = 0.008$ of all synthesis attempts. The fraction of the proteins which do not have any translation or FS mutations is given by eq.(17); for the *rpsU* gene, $\upsilon = 0.859$. The fraction of proteins which do not have any mutation at all is $\xi = 0.791$. Thus the analysis predicts that 79.1% of all *rpsU* proteins do not have transcription, translation or frameshift mutations.

Results for the dnaG gene

In Figure 2 the ribosome occupancy distribution is shown for the mRNA of the *dnaG* primase gene. This gene has 582 codons, which is considerably longer than the first example.
The 0RF occupancy drops near-continuously over the whole length of the mRNA. The probability that the ribosome shifts out-of-frame over the course of a full-length translation is $1 - P^{581}(2) = 0.0726$; this is also the fraction of all synthesis attempts that is prematurely terminated. The proteins (as a fraction of all synthesis attempts) which do not have any translation or FS mutations are $\upsilon = 0.2787$. The fraction of proteins which do not have any mutation at all is $\xi = 0.0977$. We conclude from this analysis that mutations due to mistranscription is

$\eta - \upsilon = 0.9274 - 0.2787 = 0.6487$, and mutations due to mistranslation is $\upsilon - \xi = 0.2787 - 0.0977 = 0.1810$. To summarize, of all synthesis attempts, 7.26% are terminated early due to frameshifting, 64.87% has at least one mistranscription error, 18.1% has a misreading error and 9.8% is error-free. Of course, not all mutations are lethal, but the fraction of *dnaG* primase that is error-free, is significantly lower than in the case of the *rpsU* protein.

Results for the rpoD gene

The *rpoD gene has 614 codons.* In Figure 3 the ribosome occupancy distribution is shown. The drop in 0RF occupancy is nearly linear and the probability that the ribosome occupies the ORF just before it reads the stop codon in ORF is $P^{613}(2) = 0.9335$. The fraction that is mistranscribed is $M_{Tr} = \eta - \upsilon = 0.9335 - 0.2579 = 0.6756$. The fraction that is misread during translation is $M_{Tl} = \upsilon - \xi = 0.2579 - 0.0860 = 0.1719$. Although the sequences of the *rpoD* gene and the *dnaG* gene use common and rare codons respectively, and they are of comparable size, there are not notable differences in the fractions that are misread (17.2% and 18.1%) and mistranscribed (67.6% and 64.9%). However, one cannot draw any conclusions regarding expression levels from these numbers, because expression levels depend on the rates of translation, a dynamic aspect that has not been addressed in this model.

Results for the prfB gene

The *prfB* gene has a programmed frameshift at the 26th codon position to the +1RF. There is a stop codon at this codon position in the ORF. In Figure 4 the ribosome occupation distribution is shown for the *prfB* gene. Once the frameshift has occurred, the ribosome occupies the +1RF with high probability until the 365th codon. The -1RF has a high number of stop codons that will prematurely terminate any erroneous frameshift into that frame. Another interesting finding is that there are even more stop codons present in the ORF after codon 26, than in the -1RF.

Amino Acid Composition

To demonstrate the distribution of amino acids at each codon position, the *rpsU* gene is used as an example. In Figure 5 the amino acid distribution at the first codon is shown. There are seven amino acids and their probabilities ($p(1),..p(7)$) to be incorporated in the polypeptide, vary greatly. The ordinate of Figure 5 is labeled "Fidelity' and it is defined as $\ln[10{,}000 \times p(j)+1]$, where $p(j)$ is the probability. Of the seven amino acids shown in Figure 5, methionine is the most likely amino acid to be incorporated into the polypeptide. Of the other amino acids L, V, K, T R and I, isoleucine has the highest probability of the incorrect amino acids. In Figure 6 the distribution is shown for the 40th codon position. The intended amino acid is lysine, but six other amino acids, E, T, R, I, N and Q, as well as a stop codon compete with lysine. Asparagine has the best probability to substitute the lysine.

Conclusions

An analysis of the process of encoding has been presented. All three frames are considered in the process. The subtlety of the tri-frame coding is surprising. Out-of-frame stops and pauses close to the start codon, have the function to maintain proper reading frame. In the bacterium *Escherichia coli,* efficiently translated mRNA's have an A at the start of the second codon (Looman *et al*., 1987; Sato *et al*., 2001; Stenström *et al*., 2001; Stenström and Isaksson, 2002). Therefore, efficiently translated *E.coli* N-formylmethionine initiation signals have the RNA sequence AUGA. Similarly, efficiently translated GUG and UUG initiation signals have the sequences GUGA and UUGA when the adjacent 3' nucleotide is an A, then protein translation is terminated. Alternatively, if the second codon starts with G, U or C, the codons following a frameshifting event are shown underlined as AUGG, AUGU and AUGC. All three are rare codons and the probability to frameshift again is likely to occur. The occurrence of out-of-frame stops later in the sequence plays more of a regulatory role, extending the processing time of the ribosome and thus modulating the expression levels.

The major findings of the study are summarized as follows:

- The transcription and translation processes are not deterministic.
- The consideration of all three reading frames leads to the concept of ribosome occupancy distribution.
- The serial events transcription and translation lead to a decrease in the accuracy of protein synthesis.
- The matrix V , which consists of the frameshift probabilities, and the transformation matrix *M,* which contain misreading frequencies, link (in a mathematical sense) the genetic code and intracellular aa-tRNA composition.
- Mistranscription by the RNA polymerase and mistranslation by the ribosome strongly increase the ambiguity of amino acids at each codon position. The model provides quantitative values for these occurrences.
- The degeneracy of the genetic code increases the accuracy of the synthesized protein.
- The theory gives formal expression to protein yields and mutation levels, as summarized in Table 4.
- The use of codons with high competition from near-cognates, decreases the yield and subsequently modulates the expression levels of proteins. The model is demonstrated for the genes *rpsU*, *dnaG, rpoD* and *prfB* of *E. coli*.

Acknowledgements

The authors acknowledge the financial support of the National Institutes of Health through grant R21RR022860.

References

Andersson, S.G.E., Kurland, C.G., 1990. Codon preferences in free-living organisms. Microbiol. Rev. 54, 198-210.

Antezana, M.A., Kreitman, M., 1999. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J. Mol. Evol. 49, 36-43.

Archetti, M., 2004. Selection on codon usage for error minimization at the protein level. J. Mol. Evol. 59, 400-415.

Cserzo, M., Simon, I., 1989. Regularities in the primary structure of proteins. Int. J. Peptide Res. 34, 184-195.

Dong, H., Nilsson, L., Kurland, C.G., 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J. Mol. Biol. 260, 649-63.

Farabaugh, P.J., 1997. Programmed alternative reading of the genetic code. Springer-Verlag, Heidelberg, Germany.

Fluitt, A., Pienaar, E., Viljoen, H.J., 2007. Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. Comput. Biol. Chem. 31, 335-346.

Gromadski, K.B., Rodnina, M., 2004. Kinetic determinants of high-fidelity tRNA discrimination on the ribosome. Mol. Cell 13, 191-200.

Hansen, T.M., Baranov, P.V., Ivanov, I.P., Gesteland, R.F., Atkins, J.F., 2003. Maintenance of the correct open reading frame by the ribosome. EMBO Rep. 4, 499-504.

Harger, J.W., Meskauskas, A., Dinman, J.D., 2002. An 'integrated model' of programmed ribosomal frameshifting. Trends Biochem. Sci. 27, 448-454.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 13-34.

Konigsberg, W., Godson, N., 1983. Evidence for use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia coli*. Proc. Nat. Acad. Sci. USA 80, 687-691.

Konopka, A.K., 1985. Theory of degenerate coding and informational parameters of protein coding genes. Biochimie 67, 455-468.

Kramer, E.B., Farabaugh, P.J., 2007. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. RNA 13, 87-96.

Kurland, C.G., 1979. In: Celis, J.E., Smith, J.D. (Eds.), Nonsense Mutations and tRNA Supressors, Academic Press, NY, pp. 97-108.

Lindsley, D., Gallant, J., Doneanu, C., Nonthuis, P., Caldwell, S., Fontelera, A., 2005. Spontaneous ribosome bypassing in growing cells. J. Mol. Biol. 349, 261-272.

Looman, A.C., Bodlaender, J., Comstock, L.J., Eaton, D., Jhurani, P., de Boer, J.A., van Knippenberg, P.H., 1987. Influence of the codon following the AUG initiation codon on the expression of a modified lacZ gene in *E.coli*. EMBO J. 6, 2489-2492.

Manley, J.L., 1978. Synthesis and degradation of termination and premature-termination fragments of □-galactosidase *in vitro* and *in vivo.* J. Mol. Biol. 125, 407-432.

Marquez, R., Smit, S., Knight, R., 2005. Do universal codon-usage patterns minimize the effects of mutation and translation error? Genome Biol. 6, R91.

Marquez, V., Wilson, D.N., Tate, W.P., Triana-Alonso, F., Nierhaus, K.H., 2004. Maintaining the ribosomal reading frame: the influence of the E site during translational regulation of release factor 2. Cell 118, 45-55.

Nirenberg, M., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D., Doctor, B., Hatfield, D., Levin, J., Rottman, F., Pestka, S., Wilcox, M., Anderson, F., 1966. RNA code words and protein synthesis. Cold Spring Harb. Symp. Quant. Biol. 31, 11-24.

Parker, J., 1989. Errors in reading the universal code. Microbiol. Rev. 53, 273-298.

Sato, T., Terabe, M., Watanabe, H., Gojobori, T., Hori-Takemoto, C., Miura, K-I., 2001. Codon and base biases after the initiation codon of the open reading frames in the *Escherichia coli* genome and their influence on the translation efficiency. J. Biochem. 129, 851-860.

Seligmann, H., Pollock, D.D., 2004. The ambush hypothesis: off frame stop codons arrest early accidental frameshifted transcription. DNA Cell Biol. 23, 701-705.

Sharp, P.M., Cow, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., Wright, F., 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilus, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*: a review of the considerable within species diversity. Nucleic Acids Res. 16, 8207-8211.

Sipley, J., Goldman, E., 1993. Increased ribosomal accuracy increases a programmed translational frameshift in *Escherichia coli*. Proc. Nat. Acad. Sci. USA 90, 2315-2319.

Stahl, G., McCarty, G.P., Farabaugh, P.J., 2002. Ribosome structure: revisiting the connection between translational accuracy and unconventional decoding. Trends Biochem. Sci. 27, 178-183.

Stenström, C.M., Jin, H., Major, L.L., Tate, W.P., Isaksson, L.A., 2001. Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. Gene 263, 273-284.

Stenström, C.M., Isaksson, L.A., 2002. Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3' side. Gene 288, 1-8.

Tsuchihashi, Z., 1991. Translational frameshifting in the *Escherichia coli* dnaX gene *in vitro.*" Nucleic Acids Res. 19, 2457-2462.

Weiss, R., Dunn, D., Atkins, J., Gesteland, R., 1990. The Ribosome Rubbish. In: Hill, W.E., Dahlberg, A., Garrett, R.A., Moore, P.B., Sclessinger, D., Warner, J.R. (Eds.), The Ribosome: Structure, Function, & Evolution. American Society of Microbiology, Washington DC, pp. 534-540.

Figure legends:

Figure 1: Ribosome occupancy distribution amongst the three reading frames of the mRNA of the *rpsU* gene.

Figure 2: Ribosome occupancy distribution amongst the three reading frames of the mRNA of the *dnaG* gene.

Figure 3: Ribosome occupancy distribution amongst the three reading frames of the mRNA of the *rpoD* gene.

Figure 4: Ribosome occupancy distribution amongst the three reading frames of the mRNA of the *prfB* gene.

Figure 5: Amino acid distribution at the first codon of *rpsU*

Figure 6: Amino acid distribution at the fortieth codon of *rpsU*

Table 1: Tri-Frame Stop Code: Genetically Programmed Translational Termination

(Modified from (Crick *et al*., 1961; Marshall *et al*., 1967; Nirenberg *et al*., 1966)

1st Position (5' End)	2nd Position U	2nd Position C	2nd Position G	2nd Position A	3rd Position (3' End)
U	UUU *F*	UCU *S*	UGU *C*	UAU *Y*	U
	UUC	UCC	UGC	UAC	C
	UUG *L*[b]	UCG	UGG *W*	UAG *X*[a]	G
	UUA	UCA	UGA	UAA	A
C	CUU *L*	CCU *P*	CGU *R*	CAU *H*	U
	CUC	CCC	CGC	CAC	C
	CUG *L*	CCG	CGG	CAG *Q*	G
	CUA	CCA	CGA	CAA	A
G	GUU *V*	GCU *A*	GGU *G*	GAU *D*[c]	U
	GUC	GCC	GGC	GAC	C
	GUG *V*	GCG	GGG	GAG *E*	G
	GUA	GCA	GGA	GAA	A
A	AUU *I*	ACU *T*	AGU *S*	AAU *N*	U
	AUC	ACC	AGC	AAC	C
	AUG *M*	ACG	AGG *R*	AAG *K*	G
	AUA *I*	ACA	AGA	AAA	A
	Hydrophobic			*Hydrophilic*	

L = Leucine
A = Alanine
G = Glycine
S = Serine
V = Valine
E = Glutamic Acid
K = Lysine
T = Threonine
P = Proline
D = Aspartic Acid
R = Arginine
I = Isoleucine
N = Asparagine
Q = Glutamine
F = Phenylalanine
Y = Tyrosine
H = Histidine
M = Methionine
C = Cysteine
W = Tryptophan
X = Stop

[a]Green = 3 [Frame 0] Stops at **UGA**, **UAA**, **UAG**

[b]Red = 8 [Frame +1] Stops at **NUGA** , **NUAA, NUAG (LVIM stop code)**

[c]Blue = 12 [Frame – 1] Stops at **UGAN**, **UAAN**, **UAGN (KRENDS stop code)**

Total = 23 Genetically Programmed Stops in all 3 Reading Frames.

Table 2: tRNA pool composition and arrival times (*s*).

tRNA	Amino Acid	Label	Anti-codon	Codon recognized	Molecules/ cell	Fraction	Average arrival tim
Ala1	A	1	UGC	GCU,GCA, GCG	3250	4.55	0.0014
Ala2	A	2	GGC	GCC	617	0.86	0.0073
Arg2	R	3	ACG	CGU,CGC, CGA	4752	6.65	0.0009
Arg3	R	4	CCG	CGG	639	0.89	0.007
Arg4	R	5	UCU	AGA	867	1.21	0.0052
Arg5	R	6	CCU	AGG	420	0.59	0.0107
Asn	N	7	GUU	AAC,AAU	1193	1.67	0.0038
Asp1	D	8	GUC	GAC,GAU	2396	3.35	0.0019
Cys	C	9	GCA	UGC,UGU	1587	2.22	0.0028
Gln1	Q	10	UUG	CAA	764	1.07	0.0059
Gln2	Q	11	CUG	CAG	881	1.23	0.0051
Glu2	E	12	UUC	GAA,GAG	4717	6.60	0.0009
Gly1	G	13	CCC	GGG	1068.5	1.49	0.0042
Gly2	G	14	UCC	GGA,GGG	1068.5	1.49	0.0042
Gly3	G	15	GCC	GGC,GGU	4359	6.10	0.001
His	H	16	GUG	CAC,CAU	639	0.89	0.007
Ile1	I	17	GAU	AUC,AUU	1737	2.43	0.0026
Ile2	I	18	CAU	AUA	1737	2.43	0.0026
Leu1	L	19	CAG	CUG	4470	6.25	0.001
Leu2	L	20	GAG	CUC,CUU	943	1.32	0.0048
Leu3	L	21	UAG	CUA,CUG	666	0.93	0.0067

tRNA	Amino Acid	Label	Anti-codon	Codon recognized	Molecules/ cell	Fraction	Average arrival tim
Leu4	L	22	CAA	UUG	1913	2.68	0.0023
Leu5	L	23	UAA	UUA,UUG	1031	1.44	0.0043
Lys	K	24	UUU	AAA,AAG	1924	2.69	0.0023
Met f1	M	25	CAU	AUG	1211	1.69	0.0037
Met f2	M	26	CAU	AUG	715	1.00	0.0063
Met m	M	27	CAU	AUG	706	0.99	0.0064
Phe	F	28	GAA	UUC,UUU	1037	1.45	0.0043
Pro1	P	29	CGG	CCG	900	1.26	0.005
Pro2	P	30	GGG	CCC,CCU	720	1.01	0.0063
Pro3	P	31	UGG	CCA,CCU, CCG	581	0.81	0.0077
Sec	X	32	UCA	UGA	219	0.31	0.0204
Ser1	S	33	UGA	UCA,UCU, UCG	1296	1.81	0.0035
Ser2	S	34	CGA	UCG	344	0.48	0.0131
Ser3	S	35	GCU	AGC,AGU	1408	1.97	0.0032
Ser5	S	36	GGA	UCC,UCU	764	1.07	0.0059
Thr1	T	37	GGU	ACC,ACU	104	0.15	0.0434
Thr2	T	38	CGU	ACG	541	0.76	0.0083
Thr3	T	39	GGU	ACC,ACU	1095	1.53	0.0041
Thr4	T	40	UGU	ACA,ACU, ACG	916	1.28	0.0049
Trp	W	41	CCA	UGG	943	1.32	0.0046
Tyr1	Y	42	GUA	UAC,UAU	769	1.08	0.0058
Tyr2	Y	43	GUA	UAC,UAU	1261	1.76	0.0036

tRNA	Amino Acid	Label	Anti-codon	Codon recognized	Molecules/ cell	Fraction	Average arrival tim
Val1	V	44	UAC	GUA,GUG, GUU	3840	5.37	0.0012
Val2A	V	45	GAC	GUC,GUU	630	0.88	0.0072
Val2B	V	46	GAC	GUC,GUU	635	0.89	0.0071
RF1	X	47		UAA,UAG	1200	1.68	0.0003
RF2	X	48		UAA,UGA	6000	8.39	0.0001

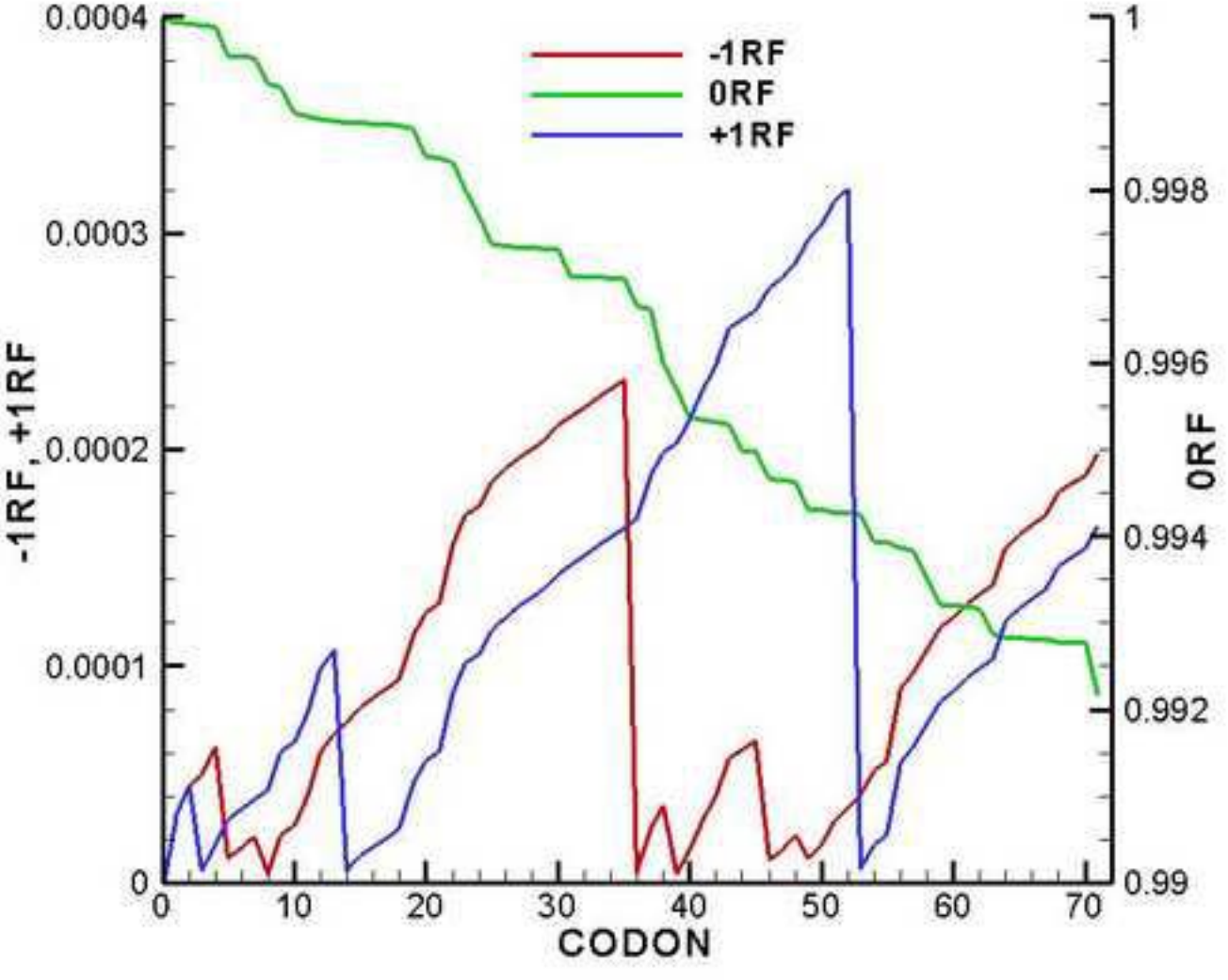
Table 3: Misread frequencies and competition measures of codons.

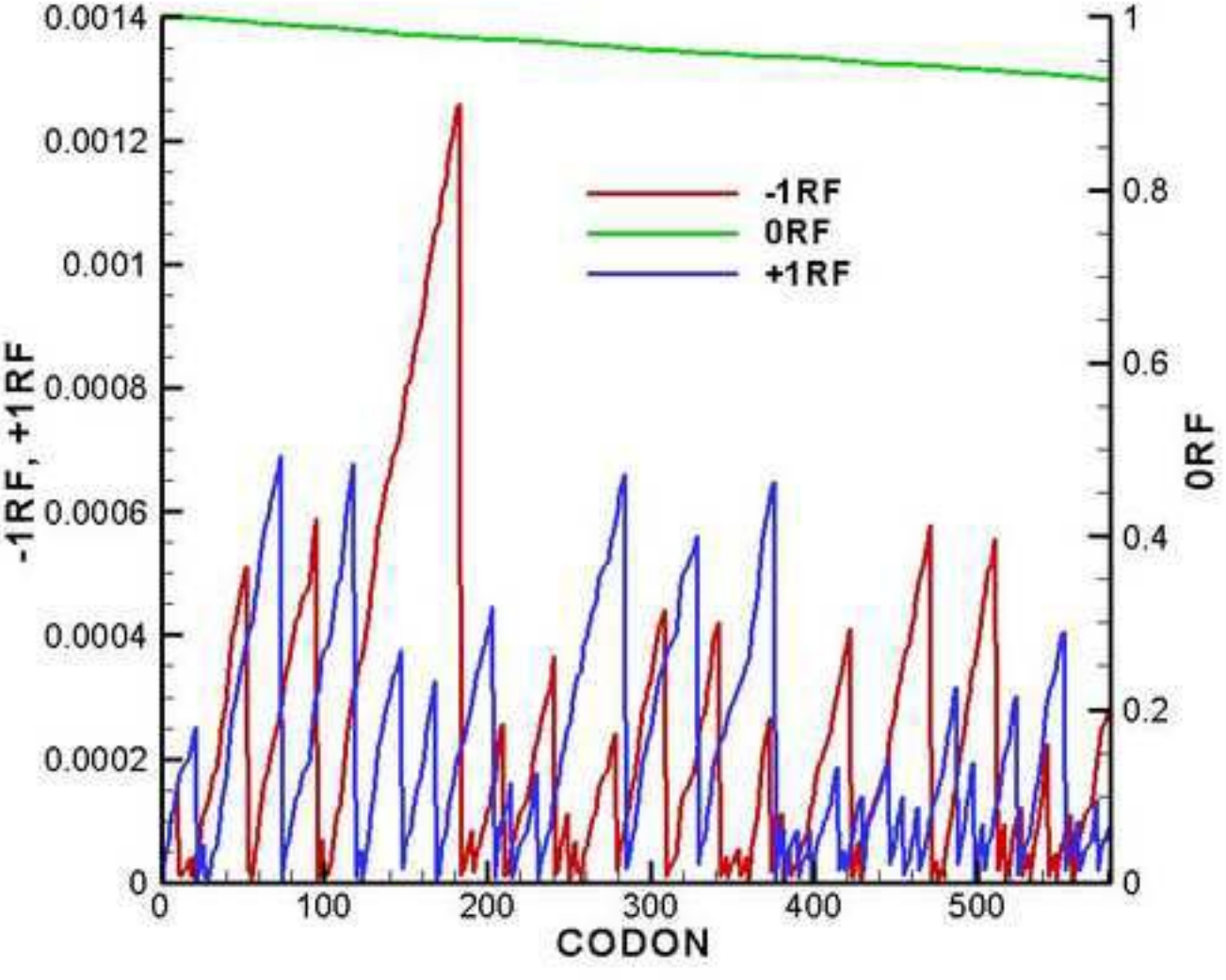
Codon	Error frequency (x 10^{-4})	Competition near-cognates / noncognates	Amino Acid of Misread tRNA	Codon	Error frequency (x 10^{-4})	Competition near-cognates /noncognates	Amino A Misread t
UUU	21	2.87 115.49	L	GUU	-	0.00 23.94	-
UUC	11, 19, 4, 13	7.07 111.29	C, L, S, Y	GUC	3, 12, 23	8.93 89.31	A, D, G
UUG	2, 1, 2	0.78 39.82	F, S, W	GUG	2	0.62 31.69	G
UUA	7, 7, 1	4.31 114.05	F, S, X	GUA	5, 8, 2	2.81 29.50	A, E, G
UCU	-	0.17 59.81	-	GCU	-	0.19 37.67	-
UCC	10, 8, 15	8.27 154.5	C, F, Y	GCC	24, 47, 13	18.40 183.24	D, G, V
UCG	7, 3	2.24 73.43	L, W	GCG	3	0.53 37.34	G
UCA	5, 1	1.85 94.31	L, X	GCA	10, 2, 7	3.25 34.61	E,G, V
UGU	3, 1	0.72 76.00	W, X	GGU	-	0.48 26.28	-
UGC	4, 2, 4, 7, 1	3.11 73.62	F, S, W, Y, X	GGC	1, 3, 2	1.42 25.34	A, D, V
UGG	12, 12, 2, 1	4.40 127.84	C, L, S, X	GGG	-	2.10 55.19	-
UGA	-	0.11 1.65	-	GGA	22, 28, 21	16.37 99.22	A, E, V
UAU	-	0.00 60.66	-	GAU	13	2.11 49.63	E
UAC	5, 4, 2	1.69 58.97	C, F, S	GAC	1, 14, 13, 4	4.80 46.94	A, E, G,
UAG	1	0.35 6.98	L	GAG	2, 1	0.69 23.29	D, G
UAA	-	0.08 1.01	-	GAA	4, 3, 1, 5	2.08 21.90	A, D, G
CUU	32	10.85 121.39	R	AUU	11	2.52 68.65	M
CUC	4, 5	6.96 125.28	H, P	AUC	4, 10, 5, 3	4.71 66.46	N, M, S
CUG	1, 1, 1	0.65 22.5	R, Q, P	AUG	5, 30, 4	9.04 167.61	R, I, T
CUA	7, 5	10.10 174.88	Q, P	AUA	4, 6, 11, 4	4.68 66.49	R, K, M
CCU	25	4.54 90.64	R	ACU	-	0.26 57.68	-
CCC	6, 8	4.29 169.59	H, L	ACC	7, 9, 8	4.81 98.17	N, I, S

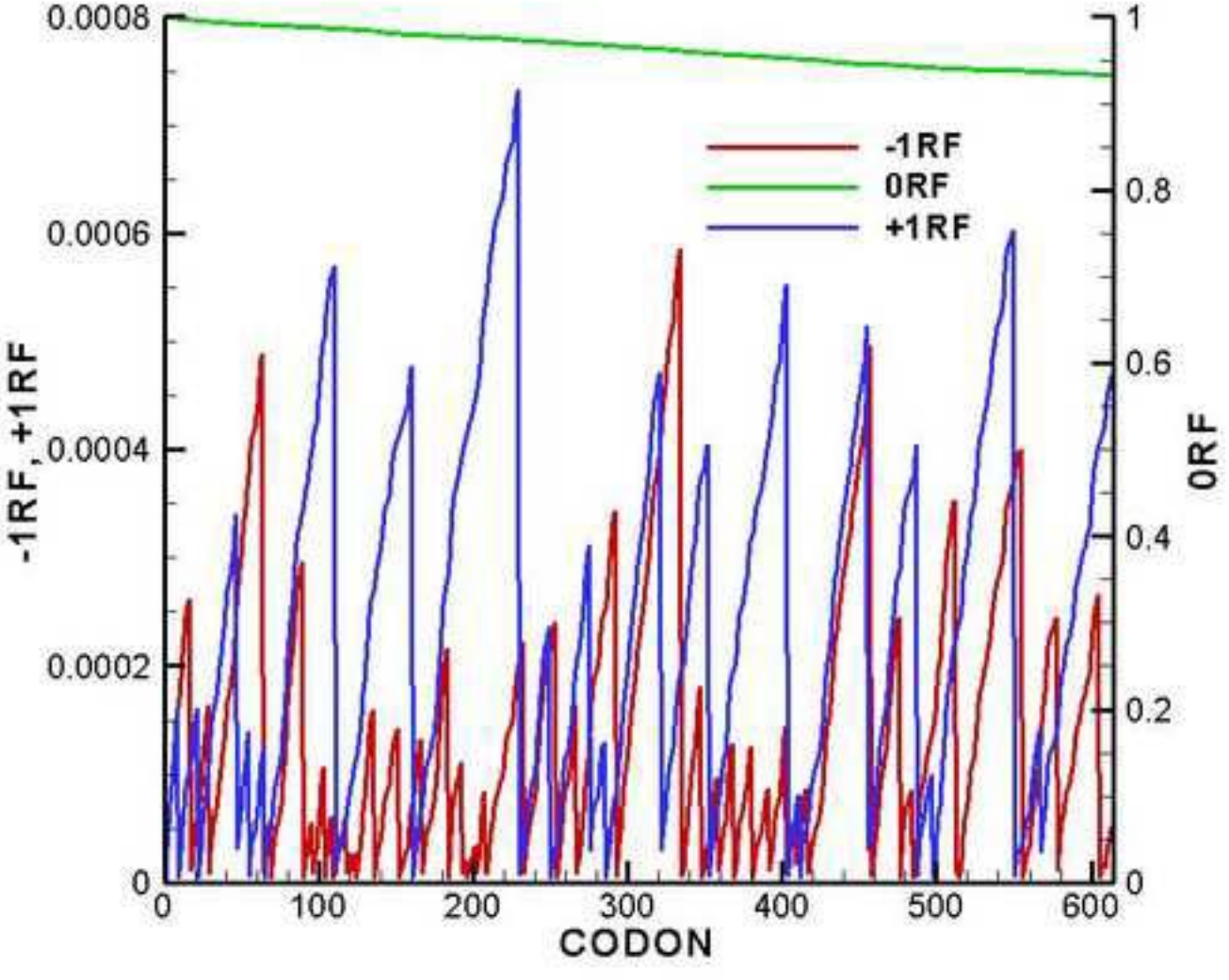
CCG	**3, 4, 18**	**4.54 78.61**	**R, Q, L**	**ACG**	**2, 8, 10**	**4.10 80.43**	**R, I, M**
CCA	**9, 8**	**5.22 207.52**	**Q, L**	**ACA**	**5, 13**	**4.97 130.04**	**R, K**
CGU	**-**	**0.13 23.85**	**-**	**AGU**	**6**	**0.91 86.91**	**R**
CGC	**1, 1**	**0.59 23.39**	**L, P**	**AGC**	**6, 6, 8, 5**	**3.84 83.98**	**R, N, I,**
CGG	**8, 46, 9**	**17.55 175.76**	**Q, L, P**	**AGG**	**24, 39, 21, 9**	**17.07 278.94**	**I, M, S,**
CGA	**1, 1, 1**	**0.53 23.45**	**Q, L, P**	**AGA**	**14, 10, 7**	**5.43 137.91**	**K, S, T**
CAU	**50, 17**	**10.34 182.97**	**R, Q**	**AAU**	**11**	**1.65 102.83**	**K**
CAC	**14, 8, 6**	**5.13 188.18**	**Q, L, P**	**AAC**	**10, 11, 8, 6**	**5.32 99.17**	**I, K, S,**
CAG	**5, 5, 33, 7**	**8.44 132.13**	**R, H, L, P**	**AAG**	**2, 3, 6, 8, 2**	**3.33 59.52**	**R, N, I,**
CAA	**5, 5, 5**	**3.65 159.13**	**H, L, P**	***AAA***	**3, 5, 2**	**1.52 61.33**	**R, N, T**

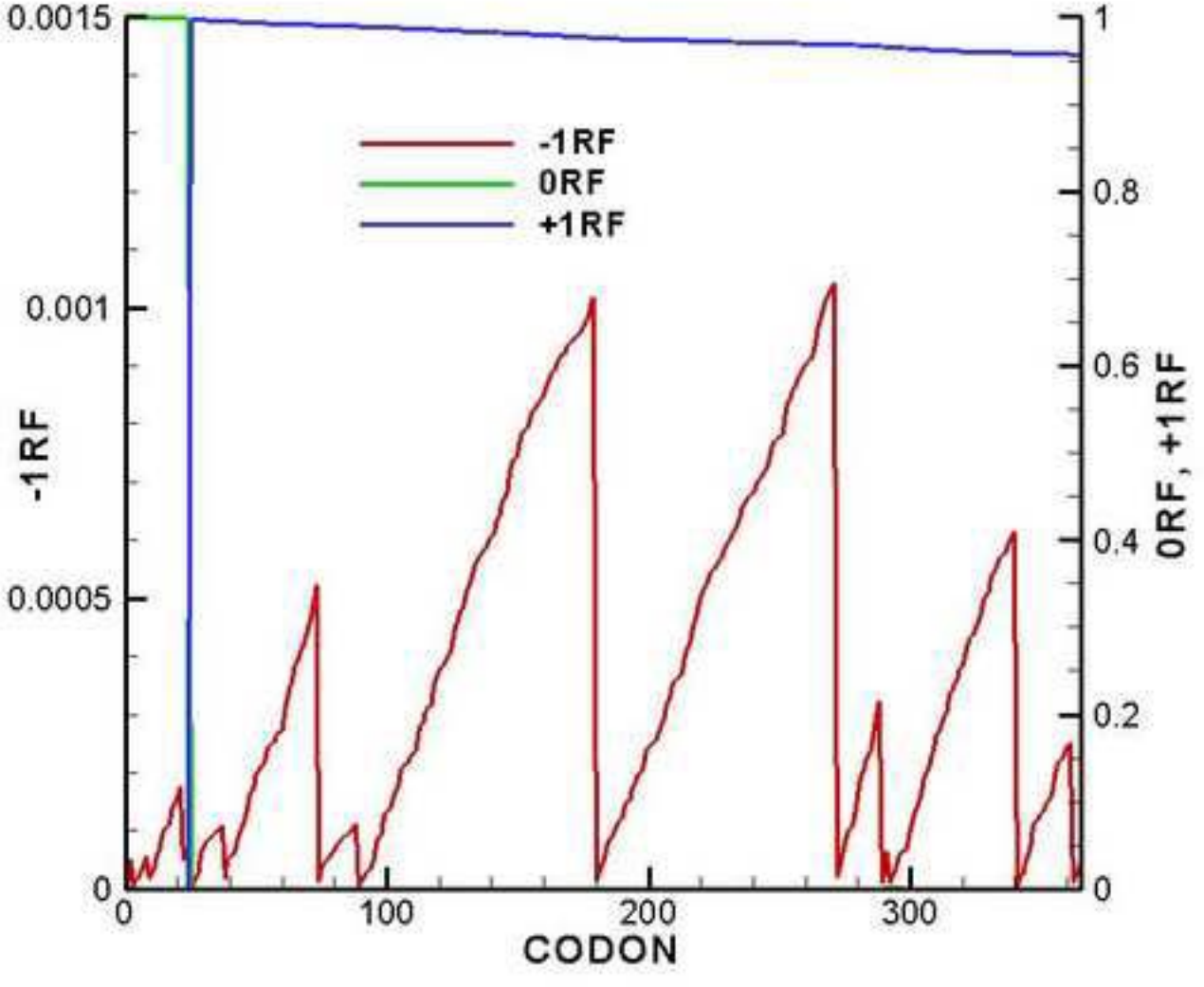
Table 4: Useful expressions of the tri-frame model

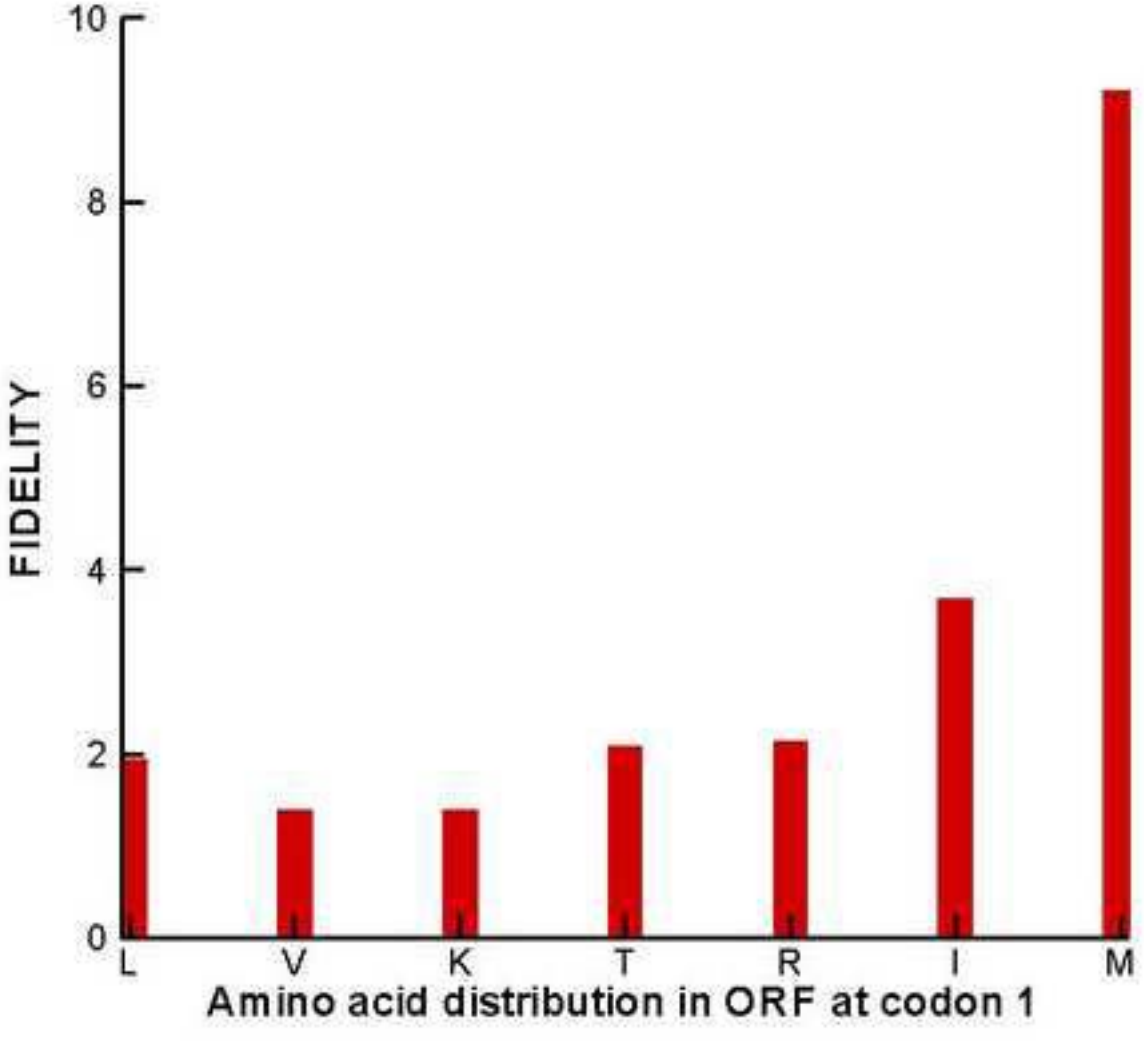
Entity	Expression
Total protein yield	$P^{N-1}(2)$, (i.e. center row of P^{N-1})
Total protein yield with no frameshift (FS) mutations	$\eta = \prod_{i=1}^{N} \overline{d}_i \cdot \overline{\alpha}$ (eq. 15)
Total protein yield with no translation or FS mutations	$\upsilon = \prod D^i_{aa}(2, a^i)\left[\overline{d}_i \cdot \overline{\alpha}\right]$ (eq. 17)
Total protein yield without any errors	$\xi = \prod S^i_{aa}(2, a^i)\left[\overline{d}_i \cdot \overline{\alpha}\right]$ (eq. 18)
Total fraction of early terminations	$\Gamma = 1 - P^{N-1}(2)$
Total mutations due to FS	$\mathrm{M}_{FS} = P^{N-1}(2) - \eta$
Total mutations due to mistranscription	$M_{Tr} = \eta - \upsilon$
Total mutations due to mistranslation	$M_{Tl} = \upsilon - \xi$
Average amino acid composition	$\left\{A^i = \left[S^i_{aa}\right]^T \times P^i, 1 \le i \le N\right\}$

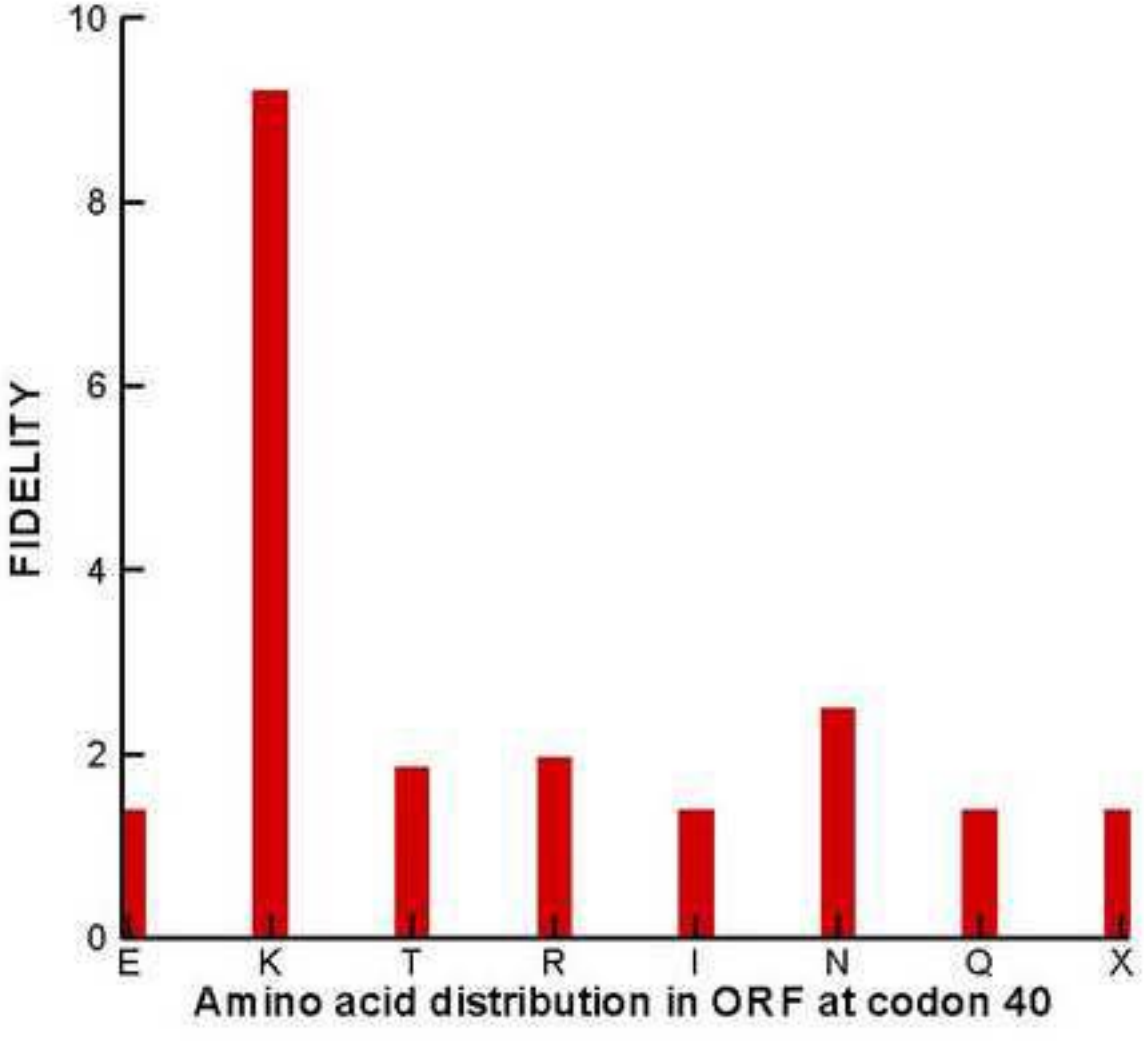

-1RF
0RF
+1RF
0.0004
0.0003
0.0002
0.0001
0
-1RF, +1RF
1
0.998
0.996
0.994
0.992
0.99
0RF
0
10
20
30
40
50
60
70
CODON


0.0014
0.0012
0.001
0.0008
0.0006
0.0004
0.0002
0
-1RF, +1RF
1
0.8
0.6
0.4
0.2
0
0RF
-1RF
0RF
+1RF
0
100
200
300
400
500
CODON


0.0008
0.0006
0.0004
0.0002
0
-1RF, +1RF
1
0.8
0.6
0.4
0.2
0
0RF
-1RF
0RF
+1RF
0
100
200
300
400
500
600
CODON


0.0015
0.001
0.0005
0
-1RF
1
0.8
0.6
0.4
0.2
0
0RF, +1RF
-1RF
0RF
+1RF
0
100
200
300
CODON


10
8
6
4
2
0
FIDELITY
L
V
K
T
R
I
M
Amino acid distribution in ORF at codon 1


FIDELITY
10
8
6
4
2
0
E
K
T
R
I
N
Q
X
Amino acid distribution in ORF at codon 40