Frameshifts and wild-type protein sequences are always similar because the genetic code is nearly optimal for frameshift tolerance

Xiaolong Wang[*1], *Quanjiang Dong*[2], *Gang Chen*[1], *Jianye Zhang*[1], *Yongqiang Liu*[1]

1. *College of Life Sciences, Ocean University of China, Qingdao, 266003, P. R. China*
2. *Qingdao Municipal Hospital, Qingdao, Shandong, 266003, P. R. China*

Abstract

Frameshift mutation yields truncated, dysfunctional product proteins, leading to loss-of-function, genetic disorders or even death. Frameshift mutations have been considered as mostly harmful and of little importance for the molecular evolution of proteins. Frameshift protein sequences, encoded by the alternative reading frames of a coding gene, have been therefore considered as meaningless. However, existing studies have shown that frameshift genes/proteins are widely exist and sometimes functional. It is puzzling how a frameshift kept its structure and functionality while its amino-acid sequence is changed substantially. We show here that the protein sequences of the frameshift remain highly conservative when compared with the wild-type protein sequence. We revealed that the similarities among the three protein sequences encoded in the three different reading frames of a coding gene are defined mainly by the genetic code. In the standard genetic code, amino acid substitutions assigned to frameshift codon substitutions are far more conservative than those assigned to random substitutions. The frameshift tolerability of the standard genetic code ranks in the top 1.0-5.0% of all possible genetic codes, showing that the genetic code is nearly optimal in terms of frameshift tolerance. In some species, the shiftability is further enhanced at gene or genome level by a biased usage of codons and codon pairs, where frameshift-tolerable codons/codon pairs are overrepresented in their genomes.

1 To whom correspondence should be addressed: *Xiaolong Wang, Ph.D., Department of Biotechnology, Ocean University of China, No. 5 Yushan Road, Qingdao, 266003, Shandong, P. R. China*, Tel: *0086-139-6969-3150*, E-mail: *Xiaolong@ouc.edu.cn*.

1. Introduction

The genetic code was deciphered in the 1960s [1-4]. The standard genetic code consists of 64 triplet codons: 61 sense codons for the twenty amino acids and the remaining three nonsense codons for stop signals. The natural genetic code has several important properties: first, the genetic code is universal for all species, with only a few variations found in some organelles or organisms, such as mitochondrion, archaea, yeast, and ciliates [5, 6]; second, the triplet codons are redundant, degenerate, and the third base is wobble (interchangeable); third, in an open reading frame, an insertion/deletion (InDel) causes a frameshift if the size of the InDel is not a multiple of three.

It has been revealed that the standard genetic code was optimized for translational error minimization [7], which is being extremely efficient at minimizing the effects of mutation or mistranslation errors [8] and is optimal for kinetic energy conservation in polypeptide chains [9]. Moreover, it was presumed that the natural genetic code resists frameshift errors by increasing the probability that a stop signal is encountered upon frameshifts because frameshifted codons for abundant amino acids overlap with stop codons [10].

Frameshifted coding genes yield truncated, non-functional, and potentially cytotoxic peptides, which lead to waste of cell energy, resources, and the activity of the biosynthetic machinery [11, 12]. Therefore, frameshift mutations have been considered as harmful and of little importance to the molecular evolution of proteins and their coding genes [13, 14]. Frameshift genes are usually considered as loss-of-function. However, frameshifted genes are sometimes functional and expressed through special mechanisms, such as translational readthrough, ribosomal frameshifting, or genetic recoding. In some species, frameshifted coding genes are tolerated by the translation systems in mitochondrial genes [15-17], e. g., a (+1) frameshift insertion is tolerated in the *nad3* gene in birds and reptiles [15]. Moreover, frameshifted overlapping genes have been found in mitochondrial genes in fruit fly and turtles [18, 19]. In *E. coli*, translational readthrough and ribosomal frameshifting were

characterized to be growth phase dependent [20]. Meanwhile, translational readthrough has been widely observed in various species, including bacteria, yeast, fruit fly, fungi, and mammals [21-27]. In addition, it has been widely observed that frameshift genes can be corrected through programmed translational (or ribosomal) frameshifting [28-31].

Moreover, it has also been reported that some frameshift genes/proteins are functional by themselves without correction. For examples: (1) in yeast, in the frameshift coding gene for mitochondrial cytochrome c oxidase subunit II (COXII), the gene is translated in an alternative frame, producing a functional COXII [32]; (2) in individuals having a frameshift mutation (138delT), alternative splicing led to the formation of a functional brain-form variant enzyme of the pseudogene cytochrome that demethylates codeine to morphine [33]. Moreover, it was reported that frameshifted genes can be retained for millions of years and enable the acquisition of new functions [34].

As is well known, proteins can be dysfunctional even by changing one single residue, it is therefore puzzling how a frameshift protein kept its tertiary structural and functional integrity while its primary sequence is being changed substantially. Here, we report that the genetic code is nearly optimal for frameshift tolerance, and protein-coding genes and genomes are also further optimized to tolerate frameshift mutations.

2. Materials and Methods

2.1 Protein and coding DNA sequences

The human/simian immunodeficiency virus (HIV/SIV) strains were derived from the seed alignment in Pfam (pf00516). The CDSs of their envelop glycoprotein (GP120) were retrieved from the HIV sequence database [35]. All available reference protein sequences and their coding DNA sequences (CDSs) in nine model organisms, including *Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Xenopus tropicalis*, *Mus musculus,* and *Homo sapiens*, were retrieved from *UCSC, Ensembl* or *NCBI* Genome Databases. Ten thousand simulated CDSs,

each with 500 random sense codons, were generated by *Recodon* 1.6.0 [36] using default settings.

2.2 Aligning and computing the similarity of the wild-type and frameshifts

Program *Frameshift-Align*.java was written to batch translate all coding sequences in their three reading frames, align their three translations, and compute their similarities. The standard genetic code was used for the translation of each CDS into three protein sequences in its three frames in the sense strand. Each internal nonsense codon was translated into a stop signal or an aa according to the *readthrough rules* (Table 1). The three translations (the wild-type and two frameshifts) were aligned by ClustalW2 using default parameters. The pairwise similarity between a frameshift and its corresponding wild-type protein sequence is given by the percent of sites in which matched amino acids are conserved (substitution score⩾0) by using the scoring matrix GON250. In the alignments, each position of a gap or a negative score was counted as a difference.

2.3 Computational analysis of frameshift codon substitutions

A protein sequence consisting of n amino acids is written as $A_1 A_2 \ldots A_i A_{i+1} \ldots A_n$, where, $A_i = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$, $i = 1 \ldots n$; its coding DNA sequence consists of n triplet codons, which is written as:

$$\boldsymbol{B_1 B_2 B_3} \mid B_4 B_5 B_6 \mid \boldsymbol{B_7 B_8 B_9} \mid \ldots \mid \boldsymbol{B_{3i+1} B_{3i+2} B_{3i+3}} \mid B_{3i+4} B_{3i+5} B_{3i+6} \mid \ldots \mid B_{3n-2} B_{3n-1} B_{3n}$$

where, $B_k = \{A, G, U, C\}$, $k = 1 \ldots 3n$. Without loss of generality, let a frameshift be caused by deleting or inserting one or two bases in the start codon:

(1) *Delete one:* $\boldsymbol{B_2 B_3} B_4 \mid B_5 B_6 \boldsymbol{B_7} \mid \ldots \mid \boldsymbol{B_{3i+2} B_{3i+3}} B_{3i+4} \mid B_{3i+5} B_{3i+6} \boldsymbol{B_{3i+7}} \mid \ldots$

(2) *Delete two:* $\boldsymbol{B_3} B_4 B_5 \mid B_6 \boldsymbol{B_7 B_8} \mid \ldots \mid \boldsymbol{B_{3i+3}} B_{3i+4} B_{3i+5} \mid B_{3i+6} \boldsymbol{B_{3i+7} B_{3i+8}} \mid \ldots$

(3) *Insert one:* $B_0 B_1 B_2 \mid \boldsymbol{B_3} B_4 B_5 \mid B_6 \boldsymbol{B_7 B_8} \mid \ldots \mid \boldsymbol{B_{3i+3}} B_{3i+4} B_{3i+5} \mid B_{3i+6} \boldsymbol{B_{3i+7} B_{3i+8}} \mid \ldots$

(4) *Insert two:* $B_{-1} B_0 \boldsymbol{B_1} \mid \boldsymbol{B_2 B_3} B_4 \mid B_5 B_6 \boldsymbol{B_7} \mid \ldots \mid \boldsymbol{B_{3i+2} B_{3i+3}} B_{3i+4} \mid B_{3i+5} B_{3i+6} \boldsymbol{B_{3i+7}} \mid \ldots$

Therefore, if a frameshift mutation occurred in the first codon, the second codon $B_4 B_5 B_6$ and its encoded amino acid A_2 has two and only two possible changes:

(1) *Forward frameshifting* (*FF*): $\boldsymbol{B_3} B_4 B_5 (\rightarrow A_{21})$

(2) *Backward frameshifting* (*BF*): $B_5 B_6 \boldsymbol{B_7}$ ($\rightarrow A_{22}$)

This continues for each of the downstream codons, resulting in two frameshifts. In either case, in every codon, all three bases are changed when compared base by base with the original codon. According to whether the encoded amino acid is changed or not, codon substitutions have been classified into two types: (1) *Synonymous substitution* (SS), (2) *Nonsynonymous substitution* (NSS). Based on the above analysis, we further classified codon substitutions into three subtypes:

(1) *Random codon substitution*: randomly change one, two, or three of the three bases of the codons, including 64×64=4096 possible codon substitutions,

(2) *Wobble codon substitution*: randomly change only the third position of the codons, including 64×4=256 possible codon substitutions,

(3) *Frameshift codon substitution*: substitutions caused by forward- or backward-frameshifting, where each has 64×4=256 possible codon substitutions.

The amino acid substitution score of a frameshift codon substitution is defined as the frameshift substitution score (FSS). Each of the 64 codons has 4 forward and 4 backward frameshift substitutions. So, there are 64×8=512 FSSs, F_{ij} and B_{ij} (i=1, …, 64; j=1, …, 4). By summing them all together, we get the total FSS of the genetic code,

$$S = \sum_{i=1}^{64} \left(\sum_{j=1}^{4} F_{ij} + \sum_{j=1}^{4} B_{ij} \right)$$

where score S is considered as the frameshift tolerability of the genetic code. Program *Frameshift-CODON*.java, was written to compute the average substitution scores for each kind of codon substitution by using a scoring matrix, BLOSSUM62 [37], PAM250 [38-40], or GON250 [41]. The substitution scores for nonsense substitutions (changing into or from a nonsense codon) are not defined in these matrices. The FSSs of nonsense substitutions are therefore calculated as zero. This is not completely reasonable but could be better than simply ignore the nonsense substitutions.

2.4 Computational analysis of alternative codon tables

Program *RandomCodes.java* was written to produce random codon tables, according to the method developed by Freeland and Hurst [8], by changing amino acids assigned to the sense codons and keeping all degenerative codons synonymous. One million of random codon tables were selected from all possible ($20! = 2.43290201\times10^{18}$) genetic codes using roulette wheel selection method. The sum of FSSs for each genetic code was computed and sorted in the ascending order and compared with that of the natural genetic code.

Program *AlternativeCodes.java* was written to produce all possible (13824) kinds of compatible alternative codon tables, proposed by Itzkovitz and Alon [10], by independently permuting the nucleotides in the three codon positions. Each alternative code has the same number of codons per amino acid and the same impact of misread errors as in the standard genetic code. The sum of FSSs for each of the compatible genetic codes was computed and sorted ascendingly and compared with that of the natural genetic code.

2.5 Computational analysis of the usage of codon and codon pairs

The usages of codons and codon pairs for the abovementioned genomes were analyzed using the method proposed in reference [42]. For each genome, the total number of codons and the number of occurrences for each codon/codon pair were counted. The observed and expected frequencies were then calculated for each of the codons and codon pairs. These calculations result in a list of 64 codons and 4096 dicodon, each with an expected (E) and observed (O) number of occurrences, usage frequency, together with a value for $\chi^2 = (O - E)^2/E$. The codons and dicodons whose O-value is greater/smaller than their E-value were identified as over-/under-represented and the weighted average FSSs were calculated for each genome.

2.6 Analysis of the frameshift substitution scores of codon pairs

For a given pair of amino acids, written as, $A_1 A_2$, where, $A_i = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$, $i = 1, 2$; its encoding codon pair is written as, $\boldsymbol{B_1 B_2 B_3}$

$| B_4 B_5 B_6$, where, $B_k = \{A, G, U, C\}$, $k = 1 \ldots 6$. There are 400 different amino acid pairs and 4096 different codon pairs.

Without loss of generality, let a frameshift be caused by inserting or deleting one base in the first codon, then, the codon pair and its encoded amino acids have two and only two types of changes:

(1) *Forward frameshifting:* $B_0 \boldsymbol{B_1 B_2} \mid \boldsymbol{B_3} B_4 B_5 (\rightarrow A_{11} A_{21})$

(2) *Backward frameshifting:* $\boldsymbol{B_2 B_3} B_4 \mid B_5 B_6 \boldsymbol{B_7} (\rightarrow A_{12} A_{22})$

Program *CODONPAIR*.java also computes the average amino acid substitution scores for each codon pair. The result of these calculations is a list of 4096 codon pairs with their corresponding FSSs, which is used to evaluate the frameshift tolerability of the codon pairs presented in a genome.

3. Results and Analysis

3.1 The definition of functional frameshift homologs

As is known, a frameshift mutation is caused by one or more InDels in a protein-coding gene whose length is not a multiple of three, and thus the reading frame is altered, fully or partially. The frameshift protein sequences encoded in the alternative reading frames of a coding gene are generally considered meaningless, as they are not only obviously different from the wild-type in the main-frame but often interrupted by many stop signals. However, we noticed that frameshifted protein sequences and their wild-type counterparts are always highly similar [43, 44], even if they are interrupted by stop signals.

We first noticed this phenomenon in an alignment of the envelop glycoproteins gene (*gp120*) of HIV/SIV [45]. As shown in Fig 1A, *gp120* underwent a series of evolutionary events, including substitution, insertion, deletion, and recombination. Especially, several frameshifting events occurred in the coding sequences (Fig 1B), but their protein sequences are highly similar. These HIV/SIV strains were originated from a common ancestor [46-48], and their GP120 proteins are surely all functional, as the infection of these viruses into

their host cells rely on these genes/proteins. In other words, *gp120* is expressed in different reading frames in different strains, but their encoded protein sequences are highly similar and all functional. Here, these frameshift GP120 are called *functional frameshift homologs*. Actually, frameshift homologs has been found widely exist in many species [49], shed light on the role of frameshift mutations in molecular evolution.

3.2 Protein sequences encoded in different reading frames are always highly similar

To validate whether or not this phenomenon is universal, we translated all coding genes of nine model organisms each into three protein sequences in all three different frames in the sense strand, aligned each of the three translations, and calculated each of their pairwise similarities. Surprisingly, in all coding sequences tested, almost all of the alignments of the three translations produce only one or a few gaps, suggesting that the three translations are always highly similar. In other words, the frameshift protein sequences and their wild-type counterparts are always highly similar in any coding sequences. For example, as shown in Fig 2, in the alignment of wild-type zebrafish VEGFAA with their frameshifts, 117/188 = 62.2% of their amino acid sites are conserved in their physiochemical properties, and the similarity of amino-acid sequences is also as high as 52.34%. This is somewhat incredible, so that we must emphasize here that this case is not a cherry picked but a common example, arbitrarily selected for the purpose of visualization. Actually, the same kind of results and figures can be reproduced by using any other coding gene sequence.

3.3 The definition of the shiftability of the protein coding genes

In a given CDS, for the three translations from the three different reading frames in the sense strand, let δ_{ij} be the similarity (the proportion of synonymous or conserved sites) between a pair of protein sequences encoded in frame *i* and frame *j,* where *i, j=1,2,3,* $i \neq j$, $\delta_{ij} = \delta_{ji}$, the average pairwise similarity among the three protein sequences is here defined as *the shiftability of the protein coding genes* (δ),

$$\delta = \frac{1}{3}(\delta_{12} + \delta_{13} + \delta_{23})$$

By analyzing all available reference CDSs in nine model organisms, we found that δ was centered approximately at 0.5 in all CDSs (Table 2 and Supplementary Dataset 1). As shown in Table 2, in all species tested, as well as in the simulated CDSs, the wild-type and the frameshift translations have a comparable amino-acid sequence similarity of 0.5. In a word, the three protein sequences encoded in the three different reading frames are always highly similar, with an average similarity of ~50%. Therefore, we propose that *protein-coding genes have a quasi-constant shiftability, equals approximately to 0.5.* In other words, in most coding genes, nearly half of their amino acids remain conserved in the frameshifts, forming the basis of frameshift tolerance of the protein-coding genes.

3.4 The readthrough rules and their impact on the computation of similarity

Translational readthrough could occur upon the activity of a suppressor tRNA with an anticodon matching a stop codon. The underlying causes for translational readthrough vary among species. The best known mechanism involves suppressor tRNAs that are expressed to correct nonsense mutations, including *amber suppressors* (*supD* [50], *supE* [51], *supF* [52]), *ochre suppressor*s (*supG* [53]), and *opal suppressors* (*supU* [52], *su9* [54]). The suppressor tRNAs frequently occur in the negative strand of a regular tRNA [55-57]. It was found that translational readthrough occurred by using these suppressor tRNAs allowing the translation of off-frame peptides [58-61]. In addition, there have been many studies reported that translational readthrough functions in *E. coli*, yeast and eukaryotes species (including human), while the readthrough rules may vary among different species [62, 63].

From known nonsense suppression tRNAs reported in *E. coli,* we summarized a set of *readthrough rules* (Table 1). In this study, these readthrough rules are used in the translating of artificial frameshift coding sequences. However, these readthrough rules are not taken as *biological laws*, but *computational methods borrowed from biology*. The only purpose of computational reading through is to obtain consecutive frameshift translations without interruption of stop signals. The frameshift protein sequences translated from the artificial frameshift CDSs do not really exist in biology but are used only for alignment and compute

the similarities of the three possible protein sequences encoded in the three different frames of a CDS. Therefore, the artificial frameshifting and reading-through operations performed on the coding sequences are conceptually different from *in-vivo translational readthrough*. In a word, computational frameshifting and reading through operations does not require or imply that these *in-silicon* readthrough rules must function in *E. coli* or any other species.

To evaluate the impact of readthrough/non-readthrough translation on the alignment of wildtype/frameshifts and their similarity calculations, readthrough and non-readthrough frameshift translations and the wild-type sequence were aligned by ClustalW. The expected proportion of nonsense codons of the total number of codons is only 3/64=4.69%, therefore, the difference of similarities computed from readthrough and non-readthrough translations is negligible in our expectation and is consistent with our observations. For example, the alignments of VEGFAA wild-type and frameshifts are fully consistent in readthrough and non-readthrough translations (Fig 2), even though a few stop signals are presented in the non-readthrough translations. The similarities of the frameshifts and the wild-type for readthrough and non-readthrough translations are close (0.5354 and 0.5573 respectively).

3.5 The genetic code was optimized for frameshift tolerance

In Table 2, the shiftability of the protein-coding genes is similar in all genes, and their standard deviations are very small in all species, suggesting that the shiftability is largely species/sequence independent, implies that the shiftability is defined mainly by the genetic code rather than the coding sequences themselves. This is also suggested by the simulated coding sequences, whose shiftability is comparable with those of the real coding genes.

As described in the method section, the average amino acid substitution scores for random, wobble, and forward and backward frameshift codon substitutions were computed respectively. As shown in Table 3 and Supplementary Dataset 2, in all 4096 random codon substitutions, only a small proportion (230/4096=5.6%) of them are synonymous, and the proportion of positive substitutions (codon substitutions with a positive substitution score) is 859/4096=20.1%. In addition, the average score of the wobble substitutions is the highest,

because most (192/256=75%) of the wobble codon substitutions are synonymous, and most (192/230=83%) of the synonymous substitutions are wobble. For frameshift substitutions, only a small proportion (28/512=5.5%) of them are synonymous (Table 4) and all the other 94.5% of them are nonsynonymous. However, the proportion of positive nonsynonymous substitutions (29.7%) is ~1.5-fold of that of the random substitutions (20.1%), and ~2-fold of that of the wobble substitutions (15.6%). *In short, in the natural (standard) genetic code, synonymous substitutions are assigned mostly to wobble codons, while frameshift codon substitutions are assigned more frequently with positive nonsynonymous substitutions.*

In addition, no matter which substitution scoring matrix (BLOSSUM62, PAM250 or GON250) is used, the average FSSs of the frameshift substitutions are always significantly higher than those of random substitutions. In GON250, for example, the average FSSs of frameshift substitutions (-1.78) is significantly higher than that of random substitutions (-10.81) (t-test $P = 2.4969 \times 10^{-10}$), suggesting that the amino acid substitutions assigned to frameshift codon substitutions are significantly more conservative than those to the random codon substitutions.

Scoring matrices are widely used to determine the similarities of amino acid sequences, to score the alignments of evolutionarily related protein sequences, and to search protein sequence databases. In all these scoring matrices, positive scores represent synonymous or similar aa substitutions, while negative scores stand for dissimilar ones. In commonly used scoring matrices, such as BLOSSUM62, PAM250, and GON250, most of the substitution scores are negative and the average percent of positive scores is only ~25%, *i. e.*, in random substitutions, the percent of positive substitutions is ~25%, and therefore, random protein sequences would have a similarity of ~25% on average.

However, as shown in Table 2, the frameshifts and the wild-type protein sequences have a similarity of ~50%, two folds of the average similarity of random protein sequences. The ~50% similarities observed among the frameshift and the wild-type protein sequences could be explained by combining the similarity derived from frameshift substitutions

(~35%) with the similarity derived from random substitutions (~25%), minus their intersection (~10%). Therefore, we concluded that the shiftability of protein-coding genes is determined by the genetic code and is largely independent on the protein/coding sequences themselves.

3.6 The natural genetic code ranks top in all possible codon tables

To further investigate the optimization of frameshift tolerability of the genetic code, we generated alternative codon tables, computed their FSSs, and compared with that of the natural genetic code. We used two strategies to creat alternative codon tables: (1) *random codon tables,* proposed by Freeland and Hurst [8], which is generated by swapping the amino acids assigned to sense codons and keeping all of the degenerative codons synonymous; (2) *compatible codon tables,* created by Itzkovitz and Alon [10], which is produced by permuting the bases in the three codon positions independently and preserving the amino acid assignment, so that each codon table has the same number of codons per each amino acid (and the same impact of misread errors) as in the standard genetic code. The total number of all possible random codon tables is an astronomical number, 20! = 2.43290201×10^{18}, while the total number of all compatible codon tables is $(4!)^3$=13824.

Using their methods, we produced one million random codon tables and all compatible codon tables, computed and sorted their FSSs (Supplementary Dataset 3). As shown in Fig 3 and Table 5, the FSSs of the natural genetic code ranks in the top ~30% of random and compatible genetic codes when they were computed using scoring matrices PAM250, but ranks in the top 1.0–5.0% of random and compatible genetic codes when computed using scoring matrices BLOSSUM62 and GON250. It is known that the new substitution scoring matrices BLOSSUM and GON are far more accurate than PAM, the oldest scoring matrices. The results given by BLOSSUM and GON are also consistent with each other, we therefore concluded that the FSS of the natural genetic codes ranks in the top 1.0–5.0% of all possible alternative codon tables, the genetic code is nearly optimal in terms of frameshift tolerance, and the shiftability of protein-coding genes is defined by the genetic code.

3.7 The shiftability is further optimized at gene/genome level

As shown in Table 2, although the shiftability of the protein-coding genes is similar in all species and all genes, and their standard deviation is very small, but many genes do have shiftability much higher than the average. In other words, although the overall shiftability is defined mainly by the genetic code, shiftability can also be adjusted at the sequence level in a certain coding gene. Functionally important coding genes, such as housekeeping genes, which are usually more conserved, may also have higher shiftability when compared with other genes. A biased usage of codons and codon pairs may contribute to sequence-level shiftability. As shown in Table 6 and Supplementary Dataset 4, in human, mouse, *Xenopus,* and yeast, the average FSSs weighted by their codon usages are significantly higher than the average FSSs of the equal usage of codons, suggesting that frameshift-tolerable codons are overrepresented in these species. However, in *E. coli, A. thaliana* and *C. elegans,* the weighted average FSSs are not higher than the unweighted average FSS, showing that frameshift-tolerable codons are not overrepresented in their genomes.

On the other hand, it has been reported that the usages of codon pairs are highly biased in virus [64], bacteria [65-67], and various species [68]. The usage of codon pairs may also influence the shiftability of coding genes. Given a pair of amino acids, $A_1 A_2$, where A_1 and A_2 have m_1 and m_2 synonymous codons, say $B_1B_2B_3$ and $B_4B_5B_6$, respectively. Therefore, the dicodon, $B_1B_2B_3|B_4B_5B_6$, has $m_1 \times m_2$ possible combinations, called *synonymous codon pairs* (SCPs). As shown in Table 7 and Supplementary Dataset 5, the usages of codon pairs are biased in all tested species. Surprisingly, in their genomes, only less than one third (up to 1660) of the 4096 possible combinations of codon pairs are over-represented. The other two third (>2400) codon pairs are under-represented or not even used, suggesting a strong selection acting on the usages of the synonymous codon pairs.

In *E. coli, C. elegans* and *A. thaliana,* because of the biased usage of codon pairs, their weighted average FSS are lower than the expected average FSS of unbiased usage of codon pairs, showing that frameshift-tolerable codon pairs are not overrepresented in these

genomes. In human, mouse, *Xenopus*, and yeast, however, their weighted average FSSs are significantly higher than the expected average FSSs, showing that frameshift-tolerable codon pairs are overrepresented in these genomes.

4. Discussion

The natural genetic code has existed since the origin of life and have been optimizing during early evolution through competition with other possible genetic codes [69-72]. The natural genetic code was optimized along several properties by selection during early evolution when compared with other possible genetic codes [73-84]. It was reported that the genetic code was optimized for translational error minimization, as the amino acids whose codons differed by a single base in the first and third positions were similar with respect to polarity and hydropathy, and the differences between amino acids, as specified by the second position, is explained by the selection to minimize the deleterious effects of translation errors during the early evolution of the genetic code [7]. In addition, it has been reported that only one in every million alternative genetic codes is more efficient than the natural genetic code in terms of minimizing the effects of point-mutations or translational errors [8]. It has been demonstrated that the natural genetic code is also nearly optimal for allowing additional information within coding genes, such as out-of-frame hidden stop codons (HSCs) and secondary structure formation (self-hybridization) [10].

There have been many hypotheses on the cause and consequence of the usages of codons and codon pairs, such as gene expression level [85-91], mRNA structure [92-98], protein abundance [98-101], and mRNA/protein stability [102-104]. Here we demonstrated that the genetic code is nearly optimal for frameshift tolerance. In addition, gene-/genome-level shiftability is achieved through a biased usage of codons/codon pairs, suggesting that genomes are further optimized for frameshift tolerance in certain organisms. This suggests that the shiftability of the protein-coding genes could be either a cause or a consequence of the usages of codons and codon-pairs. The over-represented frameshift-tolerable codons

and codon pairs may have evolutionary or survival advantages: upon a frameshift mutation, the peptide sequence of the frameshift would be more similar to the wild-type, and therefore, have better chances to remain/recover the wildtype function.

The sequence-level shiftability caused by the usages of codon/codon pairs are probably relevant to the circular code. The circular code is a set of 20 codons overrepresented in the main frame of coding genes as compared to shifted frames [105-107]. The mechanism of how the circular code maintains a reading frame is unknown [107-111], but frame detection was made possible by using the computational empirical circular code [110-115]. However, the relationship among the circular code, the biased usages of codons/codon pairs and the shiftability of coding genes remains unclear.

In this study, we discovered that the code-level shiftability of coding genes guaranteed on average over half of the sites are kept conserved in a frameshift protein when compared with the wild-type protein sequences. This is an underlying design of the natural genetic code, which forms the theoretical basis of frameshift tolerance and explains why functional frameshift homologs are widely observed [49]. Proteins and their coding genes have been evolved for billion years through frameshift and point mutations. The rate of point mutation is extremely low, and so, point mutations alter the sequence, the structure, and the function of a protein at a very slow rate. However, frameshift + point mutations provide an effective means to change protein sequences for fast evolving of novel or overlapping genes.

In reality, complete frameshifts are usually malfunctioned, and functional frameshifts are mostly partial frameshifts. Because of the shiftability, in a partially frameshifted coding genes, conservation sites include half of its frameshift codons and all of its non-frameshift codons. The protein sequence encoded by a partial frameshift is more like the wild-type than that by the complete frameshift, no matter where the partial frameshift starts and ends. This does not guarantee the proper functioning of any partial frameshifts, in general, partial frameshifts have a greater chance to recover their structure and function.

Here, we analyzed the shiftability of protein-coding genes in some model organisms. Thus, it is interesting to validate this mechanism in other genes/genomes. Conceivably, the shiftability of protein-coding genes may contribute, at least partially, to the function, repair, and evolution of frameshift proteins and their coding genes. In addition, the frameshifts are not assumed to be tolerated in the sense that they are not changed. However, if they are not removed immediately by selecting against but could be preserved in the evolutionary history, they are assumed to be repairable [116-120]. At present, we are investigating the mechanism of how a frameshift mutation is repaired [121].

5. Conclusion

Frameshift tolerance, translational readthrough, ribosomal frameshifting, and genetic recoding are widely observed in many species. Through the above analysis, we conclude that the natural genetic code is nearly optimal in terms of shiftability/frameshift tolerability. The shiftability of coding genes guarantees a half-conservation of frameshifts, endows all proteins, coding genes and organisms an inherent tolerability to frameshift mutations and suggests an adaptive advantage for the natural genetic code. As the “bottom design” for all species, the genetic code allows coding genes and genomes to tolerate frameshift mutations in both forward and backward directions, and thus, has a better fitness in the early evolution. Thanks to this ingenious property of the genetic code, the shiftability serves as an innate mechanism for cells to bear frameshift mutations of coding genes, by which the disasters of frameshift mutations were sometimes utilized as a driving force for molecular evolution.

Author Contributions

Xiaolong Wang conceived the study, coded the programs, analyzed the data, prepared the figures and tables, and wrote the paper. Yongqiang Liu analyzed some data. Quanjiang Dong, Gang Chen, and Jianye Zhang provided suggestions on the paper.

Acknowledgments

This study was funded by *National Natural Science Foundation of China* through Grant 31571369. We acknowledge Dr. Hervé Seligmanna and several other anonymous reviewers for their wonderful comments and suggestions on the earlier preprints of this article [43, 44].

Additional Information

We declare that the authors have no competing interests.

Figure Legends

Fig 1. The alignment of the coding and the protein sequences of HIV/SIV GP120. (A) The alignment of GP120 coding sequences. Highlights: showing the frameshifting events. (B) The alignment of GP120 sequences, showing that the different GP120 sequences encoded in different reading frames of *gp120* are highly similar. The alignment was aligned by ClustalW and show in GeneDoc with the bases/AAs colored by their physicochemical property.

Fig 2. The alignment of wild-type VEGFAA, readthrough or non-readthrough translation of the frameshifts. Vegfaa: wild-type VEGFAA; vegfaa-1: artificial (-1) frameshift non-readthrough translation; vegfaa-2: artificial (-2) frameshift non-readthrough translation; vegfaa-1-r: (-1) frameshift readthrough translation; vegfaa-2-r: (-2) frameshift readthrough translation;

Fig 3. The histogram of the FSSs for the genetic codes. (A) randomly chosen 1,000,000 random codon tables and (B) all 13824 compatible codon tables. FSSs were computed using scoring matrices PAM250, BLOSSUM62, and GON250, respectively. The probability densities were computed using a normal distribution function and plotted in language R.

Table 1. The *readthrough rules* derived from natural suppressor tRNAs for nonsense mutations.

Site	tRNA (AA)	Codon
supD	Ser (S)	UAG
supE	Gln (Q)	UAG
supF	Tyr (Y)	UAG
supG	Lys (K)	UAA
supU	Trp (W)	UGA

Table 2. The similarities of natural and simulated proteins and their frameshift forms.

No.	*Species*	*Number of CDSs*	*Average Similarity*					
			δ_{12}	δ_{13}	δ_{23}	δ	*MAX*	*MIN*
1	*H. sapiens*	71853	0.5217±0.0114	0.5044±0.0122	0.4825±0.0147	0.5028±0.0128	0.5948	0.4357
2	*M. musculus*	27208	0.5292±0.042	0.5058±0.0437	0.4869±0.0418	0.5073±0.0425	0.8523	0.1000*
3	*X. tropicalis*	7706	0.5190±0.0013	0.4987±0.0013	0.4855±0.0008	0.5010±0.0008	0.5962	0.4790
4	*D. rerio*	14151	0.5234±0.0007	0.5022±0.0008	0.4921±0.0005	0.5059±0.0004	0.5240	0.4784
5	*D. melanogaster*	23936	0.5162±0.0015	0.4921±0.001	0.4901±0.0013	0.4995±0.0008	0.6444	0.4667
6	*C. elegans*	29227	0.5306±0.0007	0.5035±0.0008	0.5002±0.001	0.5115±0.0006	0.6044	0.4864
7	*A. thaliana*	35378	0.5389±0.0508	0.5078±0.0481	0.5062±0.048	0.5176±0.0388	0.9540	0.2162*
8	*S. cerevisiae*	5889	0.5174±0.0011	0.4811±0.001	0.5072±0.0006	0.502±0.0007	0.5246	0.4577
9	*E. coli*	4140	0.5138±0.0019	0.4871±0.0046	0.481±0.0015	0.494±0.0012	0.7778	0.4074
10	*Simulated*	10000	0.5165±0.0282	0.4745±0.0272	0.4773±0.0263	0.4894±0.0013	0.6489	0.3539

* Very large/small similarity values were observed in a few very short or repetitive peptides.

Table 3. The amino acid substitution scores for different kinds of codon substitutions.

Codon Substitution		*ALL (Random)*	*Frameshift*		*Wobble*
			FF	***BF***	
Type of Codon Substitution	*All*	4096	256	256	256
	Unchanged (%)	64 (1.6%)	4 (1.6%)	4 (1.6%)	64 (25%)
	Changed (%)	4032 (98.4%)	252 (98.4%)	252 (98.4%)	192 (75%)
	SS (%)	230 (5.6%)	14 (5.5%)	14 (5.5%)	192 (75%)
	NSS-Positive (%)	859 (20.1%)	76 (29.7%)	76 (29.7%)	40 (15.6%)
	NSS-Negative (%)	3007 (73.4%)	166 (64.8%)	166 (64.8%)	24 (9.4%)
Average Substitution Score	*BLOSSUM62*	-1.29	-0.61	-0.65	3.77
	PAM250	-4.26	-0.84	-0.84	3.68
	GON250	-10.81	-1.78	-1.78	35.60

SS/NSS: synonymous/nonsynonymous substitution; FF/BF: forward/backward frameshift codon substitution.

Table 4. The synonymous frameshift substitutions

Forward Frameshifting					*Backward Frameshifting*				
From			*To*		*From*			*To*	
1	AAA	K	AAA	K	1	AAA	K	AAA	K
2	AAA	K	AAG	K	2	AAG	K	AAA	K
3	GGG	G	GGA	G	3	GGA	G	GGG	G
4	GGG	G	GGG	G	4	GGG	G	GGG	G
5	GGG	G	GGC	G	5	GGC	G	GGG	G
6	GGG	G	GGT	G	6	GGT	G	GGG	G
7	CCC	P	CCA	P	7	CCA	P	CCC	P
8	CCC	P	CCG	P	8	CCG	P	CCC	P
9	CCC	P	CCC	P	9	CCC	P	CCC	P
10	CCC	P	CCT	P	10	CCT	P	CCC	P
11	CTT	L	TTA	L	11	TTA	L	CTT	L
12	CTT	L	TTG	L	12	TTG	L	CTT	L
13	TTT	F	TTC	F	13	TTC	F	TTT	F
14	TTT	F	TTT	F	14	TTT	F	TTT	F

Table 5. The frameshift substitution scores of the natural and alternative genetic codes.

Genetic codes	***Scoring***	***The natural genetic code***			***FSS of the alternative genetic codes***		
(Number tested)	***Matrix***	***FSS***	***Rank***	***Rank %***	***Average***	***Max***	***Min***
Random (1,000,000)	PAM250	-344	283935	28.39%	-422.12	112.0	-1032.0
	Blossum62	-276	47340	4.73%	-411.94	-49.0	-772.0
	Gonnet250	-91.2	11675	1.17%	-323.70	166.6	-788.4
Compatible (13824)	PAM250	-344	4273	30.91%	-401.25	-140.0	-592.0
	Blossum62	-276	481	3.48%	-436.75	-250.0	-585.0
	Gonnet250	-91.2	495	3.58%	-273.61	-55.0	-481.2

Table 6. The usage of codons and their weighed average FSSs (Gon250)

NO	***Species (Codon Usage)***	***Weighted Average FSS***
1	*H. sapiens*	-9.82
2	*M. musculus*	-13.47
3	*X. tropicalis*	-12.75
4	*D. rerio*	-20.58
5	*D. melanogaster*	-19.43
6	*C. elegans*	-23.38
7	*A. thaliana*	-22.52
8	*S. cerevisiae*	-14.08
9	*E. coli*	-28.59
10	*Equal usage*	-22.27

Table 7. The usage of codon pairs and their weighed average FSSs (Gon250)

NO	*species*	*Number of codon pairs*			*Weighted Average FSS*
		Over-represented	*Under-represented*	*Absent*	
1	*H. sapiens*	1573	2523	50	-3.06
2	*M. musculus*	1505	2591	190	-3.81
3	*X. tropicalis*	1660	2436	148	-3.80
4	*D. rerio*	1493	2603	148	-5.18
5	*D. melanogaster*	1418	2678	140	-5.02
6	*C. elegans*	1469	2627	164	-6.11
7	*A. thaliana*	1566	2530	15	-6.37
8	*S. cerevisiae*	1493	2603	159	-4.27
9	*E. coli*	1389	2707	197	-6.82
10	Equal Usage	0	0	0	-5.67

References

37.

1. Nirenberg, M.W. and J.H. Matthaei, *The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides.* Proc Natl Acad Sci U S A, 1961. **47**: p. 1588-602.

2. Lacey, J.C., Jr. and K.M. Pruitt, *Origin of the genetic code.* Nature, 1969. **223**(5208): p. 799-804.

3. Crick, F.H., *The origin of the genetic code.* J Mol Biol, 1968. **38**(3): p. 367-79.

4. Reanney, D.C. and R.K. Ralph, *A speculation on the origin of the genetic code.* J Theor Biol, 1967. **15**(1): p. 41-52.

5. Jukes, T.H. and S. Osawa, *Evolutionary changes in the genetic code.* Comp Biochem Physiol B, 1993. **106**(3): p. 489-94.

6.Osawa, S., et al., *Recent evidence for evolution of the genetic code.* Microbiol Rev, 1992. **56**(1): p. 229-64.

7. Haig, D. and L.D. Hurst, *A quantitative measure of error minimization in the genetic code.* J Mol Evol, 1991. **33**(5): p. 412-7.

8. Freeland, S.J. and L.D. Hurst, *The genetic code is one in a million.* Journal of Molecular Evolution, 1998. **47**(3): p. 238-248.

9. Guilloux, A. and J.L. Jestin, *The genetic code and its optimization for kinetic energy conservation in polypeptide chains.* Biosystems, 2012. **109**(2): p. 141-4.

10. Itzkovitz, S. and U. Alon, *The genetic code is nearly optimal for allowing additional information within protein-coding sequences.* Genome Research, 2007. **17**(4): p. 405-412.

11. Seligmann, H. and D.D. Pollock, *The ambush hypothesis: hidden stop codons prevent off-frame gene reading.* DNA Cell Biol, 2004. **23**(10): p. 701-5.

12. Tse, H., et al., *Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes.* BMC Genomics, 2010. **11**: p. 491.

13. Claverie, J.M., *Detecting frame shifts by amino acid sequence comparison.* J Mol Biol, 1993. **234**(4): p. 1140-57.

14. Pellegrini, M. and T.O. Yeates, *Searching for frameshift evolutionary relationships between protein sequence families.* Proteins, 1999. **37**(2): p. 278-83.

15. Russell, R.D. and A.T. Beckenbach, *Recoding of Translation in Turtle Mitochondrial Genomes: Programmed Frameshift Mutations and Evidence of a Modified Genetic Code.* Journal of Molecular Evolution, 2008. **67**(6): p. 682-695.

16. Masuda, I., M. Matsuzaki, and K. Kita, *Extensive frameshift at all AGG and CCC codons in the mitochondrial cytochrome c oxidase subunit 1 gene of Perkinsus marinus (Alveolata; Dinoflagellata).* Nucleic Acids Research, 2010. **38**(18): p. 6186-6194.

17. Haen, K.M., W. Pett, and D.V. Lavrov, *Eight new mtDNA sequences of glass sponges reveal an extensive usage of+1 frameshifting in mitochondrial translation.* Gene, 2014. **535**(2): p. 336-344.

18. Seligmann, H., *Overlapping genetic codes for overlapping frameshifted genes in Testudines, and Lepidochelys olivacea as special case.* Comput Biol Chem, 2012. **41**: p. 18-34.

19. Seligmann, H., *An overlapping genetic code for frameshifted overlapping genes in Drosophila mitochondria: antisense antitermination tRNAs UAR insert serine.* J Theor Biol, 2012. **298**: p. 51-76.

20. Wenthzel, A.M., M. Stancek, and L.A. Isaksson, *Growth phase dependent stop codon readthrough and shift of translation reading frame in Escherichia coli.* FEBS Lett, 1998. **421**(3): p. 237-42.

21. Namy, O., et al., *Identification of stop codon readthrough genes in Saccharomyces cerevisiae.* Nucleic Acids Research, 2003. **31**(9): p. 2289-2296.

22. Loughran, G., et al., *Evidence of efficient stop codon readthrough in four mammalian genes.* Nucleic Acids Research, 2014. **42**(14): p. 8928-8938.

23. Stiebler, A.C., et al., *Ribosomal Readthrough at a Short UGA Stop Codon Context Triggers Dual Localization of Metabolic Enzymes in Fungi and Animals.* Plos Genetics, 2014. **10**(10).

24. Jungreis, I., et al., *Evidence of abundant stop codon readthrough in Drosophila and other metazoa.* Genome Research, 2011. **21**(12): p. 2096-2113.

25. Dunn, J.G., et al., *Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster.* Elife, 2013. **2**.

26. Steneberg, P. and C. Samakovlis, *A novel stop codon readthrough mechanism produces functional Headcase protein in Drosophila trachea.* Embo Reports, 2001. **2**(7): p. 593-597.

27. Williams, I., et al., *Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae.* Nucleic Acids Research, 2004. **32**(22): p. 6605-6616.

28. Chen, J., et al., *Dynamic pathways of-1 translational frameshifting.* Nature, 2014. **512**(7514): p. 328-+.

29. Dinman, J.D., *Mechanisms and implications of programmed translational frameshifting.* Wiley Interdisciplinary Reviews-Rna, 2012. **3**(5): p. 661-673.

30. Smekalova, Z. and T. Ruml, *Programmed translational frameshifting - Translation of alternative products.* Chemicke Listy, 2006. **100**(12): p. 1068-1074.

31. Farabaugh, P.J., *Programmed translational frameshifting.* Microbiological Reviews, 1996. **60**(1): p. 103-&.

32. Fox, T.D., *Five TGA "stop" codons occur within the translated sequence of the yeast mitochondrial gene for cytochrome c oxidase subunit II.* Proc Natl Acad Sci U S A, 1979. **76**(12): p. 6534-8.

33. Pai, H.V., et al., *A frameshift mutation and alternate splicing in human brain generate a functional form of the pseudogene cytochrome P4502D7 that demethylates codeine to morphine.* J Biol Chem, 2004. **279**(26): p. 27383-9.

34. Raes, J. and Y. Van de Peer, *Functional divergence of proteins through frameshift mutations.* Trends Genet, 2005. **21**(8): p. 428-31.

35. Abecasis, A.B., A.M. Vandamme, and P. Lemey, *Sequence Alignment in HIV Computational Analysis*, in *HIV Sequence Compendium*, T. Thomas, et al., Editors. 2007, Theoretical

Biology and Biophysics Group, Los Alamos National Laboratory,: Los Alamos, NM. LA-UR 07-4826. p. 2-16.

36. Arenas, M. and D. Posada, *Recodon: coalescent simulation of coding DNA sequences with recombination, migration and demography.* BMC Bioinformatics, 2007. **8**: p. 458.

37. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.

38. Dayhoff, M.O., *The origin and evolution of protein superfamilies.* Fed Proc, 1976. **35**(10): p. 2132-8.

39. Dayhoff, M.O., *Computer analysis of protein sequences.* Fed Proc, 1974. **33**(12): p. 2314-6.

40. Dayhoff, M.O., *Computer analysis of protein evolution.* Sci Am, 1969. **221**(1): p. 86-95.

41. Schneider, A., G.M. Cannarozzi, and G.H. Gonnet, *Empirical codon substitution matrix.* BMC Bioinformatics, 2005. **6**: p. 134.

42. Gutman, G.A. and G.W. Hatfield, *Nonrandom utilization of codon pairs in Escherichia coli.* Proceedings of the National Academy of Sciences of the United States of America, 1989. **86**(10): p. 3699-3703.

43. Wang, X.W., X.; Chen, G.; Zhang, J.; Liu, Y.; Yang C. , *The shiftability of protein coding genes: the genetic code was optimized for frameshift tolerating.* PeerJ PrePrints 2015. **3** p. e806v1.

44. Wang, X., et al., *Why are frameshift homologs widespread within and across species?* bioRxiv, 2016.

45. Wang, X. and C. Yang, *CAUSA 2.0: accurate and consistent evolutionary analysis of proteins using codon and amino acid unified sequence alignments.* PeerJ PrePrints **3**.

46. Holmes, E.C., *On the origin and evolution of the human immunodeficiency virus (HIV).* Biol Rev Camb Philos Soc, 2001. **76**(2): p. 239-54.

47. Rambaut, A., et al., *Human immunodeficiency virus. Phylogeny and the origin of HIV-1.* Nature, 2001. **410**(6832): p. 1047-8.

48. Paraskevis, D., et al., *Analysis of the evolutionary relationships of HIV-1 and SIVcpz sequences using bayesian inference: implications for the origin of HIV-1.* Mol Biol Evol, 2003. **20**(12): p. 1986-96.

49. Antonov, I., et al., *Identification of the nature of reading frame transitions observed in prokaryotic genomes.* Nucleic Acids Res, 2013. **41**(13): p. 6514-30.

50. Nagano, T., Y. Kikuchi, and Y. Kamio, *High expression of the second lysine decarboxylase gene, ldc, in Escherichia coli WC196 due to the recognition of the stop codon (TAG), at a position which corresponds to the 33th amino acid residue of sigma(38), as a serine residue by the amber suppressor, supD.* Bioscience Biotechnology and Biochemistry, 2000. **64**(9): p. 2012-2017.

51. Kuriki, Y., *Temperature-Sensitive Amber Suppression of Ompf'-'Lacz Fused Gene-Expression in a Supe Mutant of Escherichia-Coli K12.* Fems Microbiology Letters, 1993. **107**(1): p. 71-76.

52. Johnston, H.M. and J.R. Roth, *UGA suppressor that maps within a cluster of ribosomal protein genes.* J Bacteriol, 1980. **144**(1): p. 300-5.

53. Prather, N.E., B.H. Mims, and E.J. Murgola, *supG and supL in Escherichia coli code for mutant lysine tRNAs+.* Nucleic Acids Res, 1983. **11**(23): p. 8283-6.

54. Chan, T.S. and A. Garen, *Amino acid substitutions resulting from suppression of nonsense mutations. V. Tryptophan insertion by the Su9 gene, a suppressor of the UGA nonsense triplet.* J Mol Biol, 1970. **49**(1): p. 231-4.

55. Seligmann, H., *Undetected antisense tRNAs in mitochondrial genomes?* Biol Direct, 2010. **5**: p. 39.

56. Seligmann, H., *Avoidance of antisense, antiterminator tRNA anticodons in vertebrate mitochondria.* Biosystems, 2010. **101**(1): p. 42-50.

57. Seligmann, H., *Pathogenic mutations in antisense mitochondrial tRNAs.* J Theor Biol, 2011. **269**(1): p. 287-96.

58. Seligmann, H., *Overlapping genetic codes for overlapping frameshifted genes in Testudines, and Lepidochelys olivacea as special case.* Computational Biology and Chemistry, 2012. **41**: p. 18-34.

59. Seligmann, H., *An overlapping genetic code for frameshifted overlapping genes in Drosophila mitochondria: Antisense antitermination tRNAs UAR insert serine.* Journal of Theoretical Biology, 2012. **298**: p. 51-76.

60. Seligmann, H., *Two genetic codes, one genome: Frameshifted primate mitochondrial genes code for additional proteins in presence of antisense antitermination tRNAs.* Biosystems, 2011. **105**(3): p. 271-285.

61. Faure, E., et al., *Probable presence of an ubiquitous cryptic mitochondrial gene on the antisense strand of the cytochrome oxidase I gene.* Biol Direct, 2011. **6**: p. 56.

62. Dabrowski, M., Z. Bukowy-Bieryllo, and E. Zietkiewicz, *Translational readthrough potential of natural termination codons in eucaryotes--The impact of RNA sequence.* RNA Biol, 2015. **12**(9): p. 950-8.

63. Schueren, F. and S. Thoms, *Functional Translational Readthrough: A Systems Biology Perspective.* PLoS Genet, 2016. **12**(8): p. e1006196.

64. Bizinoto, M.C., et al., *Codon pairs of the HIV-1 vif gene correlate with CD4+T cell count.* Bmc Infectious Diseases, 2013. **13**.

65. Wu, X.M., et al., *Computational identification of rare codons of Escherichia coli based on codon pairs preference.* Bmc Bioinformatics, 2010. **11**.

66. Boycheva, S., G. Chkodrov, and I. Ivanov, *Codon pairs in the genome of Escherichia coli.* Bioinformatics, 2003. **19**(8): p. 987-998.

67. Boycheva, S.S. and I.G. Ivanov, *Missing codon pairs in the genome of Escherichia coli.* Biotechnology & Biotechnological Equipment, 2002. **16**(1): p. 142-144.

68. Tats, A., T. Tenson, and M. Remm, *Preferred and avoided codon pairs in three domains of life.* Bmc Genomics, 2008. **9**.

69. Morgens, D.W. and A.R. Cavalcanti, *An alternative look at code evolution: using non-canonical codes to evaluate adaptive and historic models for the origin of the genetic code.* J Mol Evol, 2013. **76**(1-2): p. 71-80.

70. Santos, M.A., et al., *Driving change: the evolution of alternative genetic codes.* Trends Genet, 2004. **20**(2): p. 95-102.

71. Ardell, D.H. and G. Sella, *On the evolution of redundancy in genetic codes.* J Mol Evol, 2001. **53**(4-5): p. 269-81.

72. Maeshiro, T. and M. Kimura, *The role of robustness and changeability on the origin and evolution of genetic codes.* Proc Natl Acad Sci U S A, 1998. **95**(9): p. 5088-93.

73. Trifonov, E.N., *Evolution of the Genetic Code and the Earliest Proteins.* Origins of Life and Evolution of Biospheres, 2009. **39**(3-4): p. 184-184.

74. Koonin, E.V. and A.S. Novozhilov, *Origin and Evolution of the Genetic Code: The Universal Enigma.* Iubmb Life, 2009. **61**(2): p. 99-111.

75. Archetti, M. and M. Di Giulio, *The evolution of the genetic code took place in an anaerobic environment.* Journal of Theoretical Biology, 2007. **245**(1): p. 169-174.

76. Wiltschi, B. and N. Budisa, *Natural history and experimental evolution of the genetic code.* Applied Microbiology and Biotechnology, 2007. **74**(4): p. 739-753.

77. Travers, A., *The evolution of the genetic code revisited.* Origins of Life and Evolution of the Biosphere, 2006. **36**(5-6): p. 549-555.

78. Knight, R.D. and L.F. Landweber, *The early evolution of the genetic code.* Cell, 2000. **101**(6): p. 569-572.

79. Jimenez-Montano, M.A., *Protein evolution drives the evolution of the genetic code and vice versa.* Biosystems, 1999. **54**(1-2): p. 47-64.

80. Davis, B.K., *Evolution of the genetic code.* Progress in Biophysics & Molecular Biology, 1999. **72**(2): p. 157-243.

81. JimenezSanchez, A., *On the origin and evolution of the genetic code.* Journal of Molecular Evolution, 1995. **41**(6): p. 712-716.

82. Beland, P. and T.F.H. Allen, *The Origin and Evolution of the Genetic-Code.* Journal of Theoretical Biology, 1994. **170**(4): p. 359-365.

83. Baumann, U. and J. Oro, *3 Stages in the Evolution of the Genetic-Code.* Biosystems, 1993. **29**(2-3): p. 133-141.

84. Osawa, S., et al., *Recent-Evidence for Evolution of the Genetic-Code.* Microbiological Reviews, 1992. **56**(1): p. 229-264.

85. Coghlan, A. and K.H. Wolfe, *Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae.* Yeast, 2000. **16**(12): p. 1131-45.

86. Willie, E. and J. Majewski, *Evidence for codon bias selection at the pre-mRNA level in eukaryotes.* Trends Genet, 2004. **20**(11): p. 534-8.

87. Goetz, R.M. and A. Fuglsang, *Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli.* Biochem Biophys Res Commun, 2005. **327**(1): p. 4-7.

88. Roymondal, U., S. Das, and S. Sahoo, *Predicting gene expression level from relative codon usage bias: an application to Escherichia coli genome.* DNA Res, 2009. **16**(1): p. 13-30.

89. Herbeck, J.T., D.P. Wall, and J.J. Wernegreen, *Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont Wigglesworthia.* Microbiology, 2003. **149**(Pt 9): p. 2585-96.

90. Li, H. and L. Luo, *The relation between codon usage, base correlation and gene expression level in Escherichia coli and yeast.* J Theor Biol, 1996. **181**(2): p. 111-24.

91. Paul, P., A.K. Malakar, and S. Chakraborty, *Codon usage and amino acid usage influence genes expression level.* Genetica, 2017.

92. Shen, X., S. Chen, and G. Li, *Role for gene sequence, codon bias and mRNA folding energy in modulating structural symmetry of proteins.* Conf Proc IEEE Eng Med Biol Soc, 2013. **2013**: p. 596-9.

93. Pop, C., et al., *Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation.* Mol Syst Biol, 2014. **10**: p. 770.

94. Carlini, D.B., Y. Chen, and W. Stephan, *The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr.* Genetics, 2001. **159**(2): p. 623-33.

95. Griswold, K.E., et al., *Effects of codon usage versus putative 5'-mRNA structure on the expression of Fusarium solani cutinase in the Escherichia coli cytoplasm.* Protein Expr Purif, 2003. **27**(1): p. 134-42.

96. Gambari, R., C. Nastruzzi, and R. Barbieri, *Codon usage and secondary structure of the rabbit alpha-globin mRNA: a hypothesis.* Biomed Biochim Acta, 1990. **49**(2-3): p. S88-93.

97. Zama, M., *Codon usage and secondary structure of mRNA.* Nucleic Acids Symp Ser, 1990(22): p. 93-4.

98. Subramanian, A. and R.R. Sarkar, *Comparison of codon usage bias across Leishmania and Trypanosomatids to understand mRNA secondary structure, relative protein abundance and pathway functions.* Genomics, 2015. **106**(4): p. 232-41.

99. Klumpp, S., J. Dong, and T. Hwa, *On ribosome load, codon bias and protein abundance.* PLoS One, 2012. **7**(11): p. e48542.

100. Zhou, J.H., et al., *The effects of the synonymous codon usage and tRNA abundance on protein folding of the 3C protease of foot-and-mouth disease virus.* Infect Genet Evol, 2013. **16**: p. 270-4.

101. McHardy, A.C., et al., *Comparing expression level-dependent features in codon usage with protein abundance: an analysis of 'predictive proteomics'.* Proteomics, 2004. **4**(1): p. 46-58.

102. Mukhopadhyay, P., S. Basak, and T.C. Ghosh, *Synonymous codon usage in different protein secondary structural classes of human genes: implication for increased non-randomness of GC3 rich genes towards protein stability.* J Biosci, 2007. **32**(5): p. 947-63.

103. Stenoien, H.K. and W. Stephan, *Global mRNA stability is not associated with levels of gene expression in Drosophila melanogaster but shows a negative correlation with codon bias.* J Mol Evol, 2005. **61**(3): p. 306-14.

104. Mishima, Y. and Y. Tomari, *Codon Usage and 3' UTR Length Determine Maternal mRNA Stability in Zebrafish.* Mol Cell, 2016. **61**(6): p. 874-85.

105. Arques, D.G., J.P. Fallot, and C.J. Michel, *An evolutionary analytical model of a complementary circular code simulating the protein coding genes, the 5' and 3' regions.* Bull Math Biol, 1998. **60**(1): p. 163-94.

106. Arques, D.G. and C.J. Michel, *A complementary circular code in the protein coding genes.* J Theor Biol, 1996. **182**(1): p. 45-58.

107. El Soufi, K. and C.J. Michel, *Circular code motifs in genomes of eukaryotes.* J Theor Biol, 2016. **408**: p. 198-212.

108. Michel, C.J., *Circular code motifs in transfer RNAs.* Comput Biol Chem, 2013. **45**: p. 17-29.

109. Michel, C.J., *Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes.* Comput Biol Chem, 2012. **37**: p. 24-37.

110. Michel, C.J., *An extended genetic scale of reading frame coding.* J Theor Biol, 2015. **365**: p. 164-74.

111. Michel, C.J., *A genetic scale of reading frame coding.* J Theor Biol, 2014. **355**: p. 83-94.

112. Ahmed, A., G. Frey, and C.J. Michel, *Frameshift signals in genes associated with the circular code.* In Silico Biol, 2007. **7**(2): p. 155-68.

113. Ahmed, A., G. Frey, and C.J. Michel, *Essential molecular functions associated with the circular code evolution.* J Theor Biol, 2010. **264**(2): p. 613-22.

114. Michel, C.J., *The maximal C(3) self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses.* J Theor Biol, 2015. **380**: p. 156-77.

115. Michel, C.J., M. Pellegrini, and G. Pirillo, *Maximal dinucleotide and trinucleotide circular codes.* J Theor Biol, 2016. **389**: p. 40-6.

116. Maletzki, C., et al., *Frameshift mutational target gene analysis identifies similarities and differences in constitutional mismatch repair-deficiency and Lynch syndrome.* Mol Carcinog, 2017. **56**(7): p. 1753-1764.

117. Woerner, S.M., et al., *Detection of coding microsatellite frameshift mutations in DNA mismatch repair-deficient mouse intestinal tumors.* Mol Carcinog, 2015. **54**(11): p. 1376-86.

118. Greene, C.N. and S. Jinks-Robertson, *Spontaneous frameshift mutations in Saccharomyces cerevisiae: accumulation during DNA replication and removal by proofreading and mismatch repair activities.* Genetics, 2001. **159**(1): p. 65-75.

119. Harfe, B.D. and S. Jinks-Robertson, *Removal of frameshift intermediates by mismatch repair proteins in Saccharomyces cerevisiae.* Mol Cell Biol, 1999. **19**(7): p. 4766-73.

120. Dohet, C., R. Wagner, and M. Radman, *Methyl-directed repair of frameshift mutations in heteroduplex DNA.* Proc Natl Acad Sci U S A, 1986. **83**(10): p. 3395-7.

121. Wang, X., et al., *Premature termination codons signaled targeted repair of frameshift mutation by endogenous RNA-directed gene editing.* bioRxiv, 2016.

```
                    *         20          *         40
HV1J3 : -----------ATGAGAGTGAAGGGGATCAGGAAGAA--ITA :   29
SIVCZ : -----------ATGAAAGTAATGGAGAAGAAGAAGAG--AGA :   29
SIVGB : ATGTCTACAGGAAACGTGTACCAGGAACTAATAAGAAGATAC :   42

                  *         60          *         80
HV1J3 : TCAGCACTTGTGGAGATGGGGCACGATGCTCCTTGGGATATT :   71
SIVCZ : CTGGAACAGCTTATCCATAATTACAATCATAACAATCATTTT :   71
SIVGB : CTGGTAGTGGTGAAGAAGCTATACGAAGGTAAGTATGAAGTG :   84

               *        100          *        120
HV1J3 : GATGATCTGTAGTGCTGCAGAACAATTGTGGGTCACAGTC-- :  111
SIVCZ : GCTAACCCCATGTTTGACCTCTGAGTTATGGGTAACAGTA-- :  111
SIVGB : TCCAGGTCTTTTTCTTATACTATGTTTA-GCCTACTAGTAGG :  125

            *        140          *        160
HV1J3 : TATTATGGGGTACCTGTGTGGAAAGAAGCAGCCACCACTCTA :  153
SIVCZ : TATTATGGAGTACCTGTTTGGCATGATGCTGACCCGGTACTC :  153
SIVGB : TATTATAGGAAAACAATATGTGACAGT-CTTCTATGGAGTAC :  166

          *        180          *        200          *
HV1J3 : TTTTGTGCATCAGATGCTAAAGCATAT---------GATACA :  186
SIVCZ : TTTTGTGCCTCAGACGCTAAGGCACAT---------AGTACA :  186
SIVGB : CAGTATGGAA-GGAAGCTAAAACACATTTGATTTGTGCTACA :  207

                 220          *        240          *
HV1J3 : GAGGTACATAATGTTTGGGCCACACATGCCTGTGTACCCACA :  228
SIVCZ : GAGGCTCATAATATTTGGGCCACACAGGCATGTGTACCTACA :  228
SIVGB : GATAATTCAAGTCTCTGGGTAACCACTAATTGCATACCTTCA :  249

               260          *        280          *
HV1J3 : GACCCCAACCCACAAGAAGTAGTATTGGAAAATGTGACAGAA :  270
SIVCZ : GATCCCAGTCCTCAGGAAGTATTTCTTCCAAATGTAATAGAA :  270
SIVGB : TTGCCAGATTATGATGAGGTAGAAATTCCTGATATAAAGGAA :  291

             300          *        320          *
HV1J3 : AAATTTAA------CATGTGGAAAAATAACATGGTAGAACAG :  306
SIVCZ : TCATTTAA------CATGTGGAAAAATAATATGGTGGACCAA :  306
SIVGB : AATTTTACAGGACTTATAAGGGAAAATCAGATAGTTTATCAA :  333
```

Fig 1 (A). Alignment of coding sequences of HIV/SIV GP120

```
                    *         20          *         40
HV1J3 : --------------MRVKGIRKNYQHLWRWGTMLLGILMICSA :  29
SIVCZ : --------------MKVMEKKKRDWNSLSIITIITIILLTPCL :  29
SIVGB : MSTGNVYQELIRRYLVVVKKLYEGKYEVSRSFSYTMFSLLVGI :  43
                      6 V  k k      s  t  t il6

              *         60          *         80
HV1J3 : AEQLWVTVYYGVPVWKEAATTLFCASDAKAYDTEVHNVWATHA :  72
SIVCZ : TSELWVTVYYGVPVWHDADPVLFCASDAKAHSTEAHNIWATQA :  72
SIVGB : IGKQYVTVFYGVPVWKEAKTHLICATDNSS-------LWVTTN :  79
           l5VTV5YGVPVWkeA t LfCA3Daka  te hn6WaT a

            *       100          *       120
HV1J3 : CVPTDPNPQEVVLENVTEKFN--MWKNNMVEQMHEDIISLWDQ : 113
SIVCZ : CVPTDPSPQEVFLPNVIESFN--MWKNNMVDQMHEDIISLWDQ : 113
SIVGB : CIPSLPDYDEVEIPDIKENFTGLIRENQIVYQAWHAMGSMLDT : 122
        C6P3dP pqEV 6p16 E Fn  6wkNn6V Qmhed6iS6wDq

        *       140          *       160          *
HV1J3 : SLKPCVKLTPLCVTLNCIDWGNDTSPNATNTTSSGGEKMEKGE : 156
SIVCZ : SLKPCVKLTPLCVTLQCSKANFSQAKNLTNQTSS-----PPLE : 151
SIVGB : ILKPCVKINPYCVKMQCQETENVSATTAKPITTPTTTSTVASS : 165
        sLKPCVK6tPlCVt6qC    n  a natn T3s        e

             180          *       200          *
HV1J3 : MKNCSFNITTSIRDKVQKEHALFY------KHDVVPINNSTKD : 193
SIVCZ : MKNCSFNVTTELRDKKKQVYSLFY------VEDVVNLG----- : 183
SIVGB : TEIYLDVDKNNTEEKVERNHVCRYNITGLCRDSKEEIVTNFRG : 208
        mkncsfn tt  rdKv   h lfY        dvv 6

          220          *       240          *       2
HV1J3 : NIKNDNSTRYRLISCNTSVITQACPKISFEPIPIHYCAPAGFA : 236
SIVCZ : ---NENNT-YRIINCNTTAITQACPKTSFEPIPIHYCAPAGFA : 222
SIVGB : DDVKCENNTCYMNHCNESVNTEDCQKG-LLIRCILGCVPPGYV : 250
           n nnt yr6i CNt3viT2aCpK sfepipIhyCaPaG5a

        60          *       280          *       300
HV1J3 : IIKCNDKKFNGTGPCTNVSTVQCTHGIKPVVSTQLLLNGSLAE : 279
SIVCZ : ILKCNDKDFSGKGKCTNVSTVHCTHGIKPVVTTQLLINGSLAE : 265
SIVGB : MLRYN-EKLNNNKLCSNISAVQCTQHLVATVSSFFGFNGTMHK : 292
        664cNdkkfng g C3N6StVqCThg6kpvV33qll NG36ae

              *       320          *       340
HV1J3 : EEVVIRSENFTDNAK-------TIIVQLKEPVVINCTRPSKTT : 315
SIVCZ : GNITVRVENKSKNTD-------VWIVQLVEAVSLNCHRPGNNT : 301
SIVGB : EGELIPIDDKYRGPEEFHQRKFVYKVPGKYGLKIECHRKGNRS : 335
        e   6r e1k  n          v iVqlke 6 6nChRpgn 3
```

Fig 1 (B). Alignment of protein sequences of HIV/SIV GP120

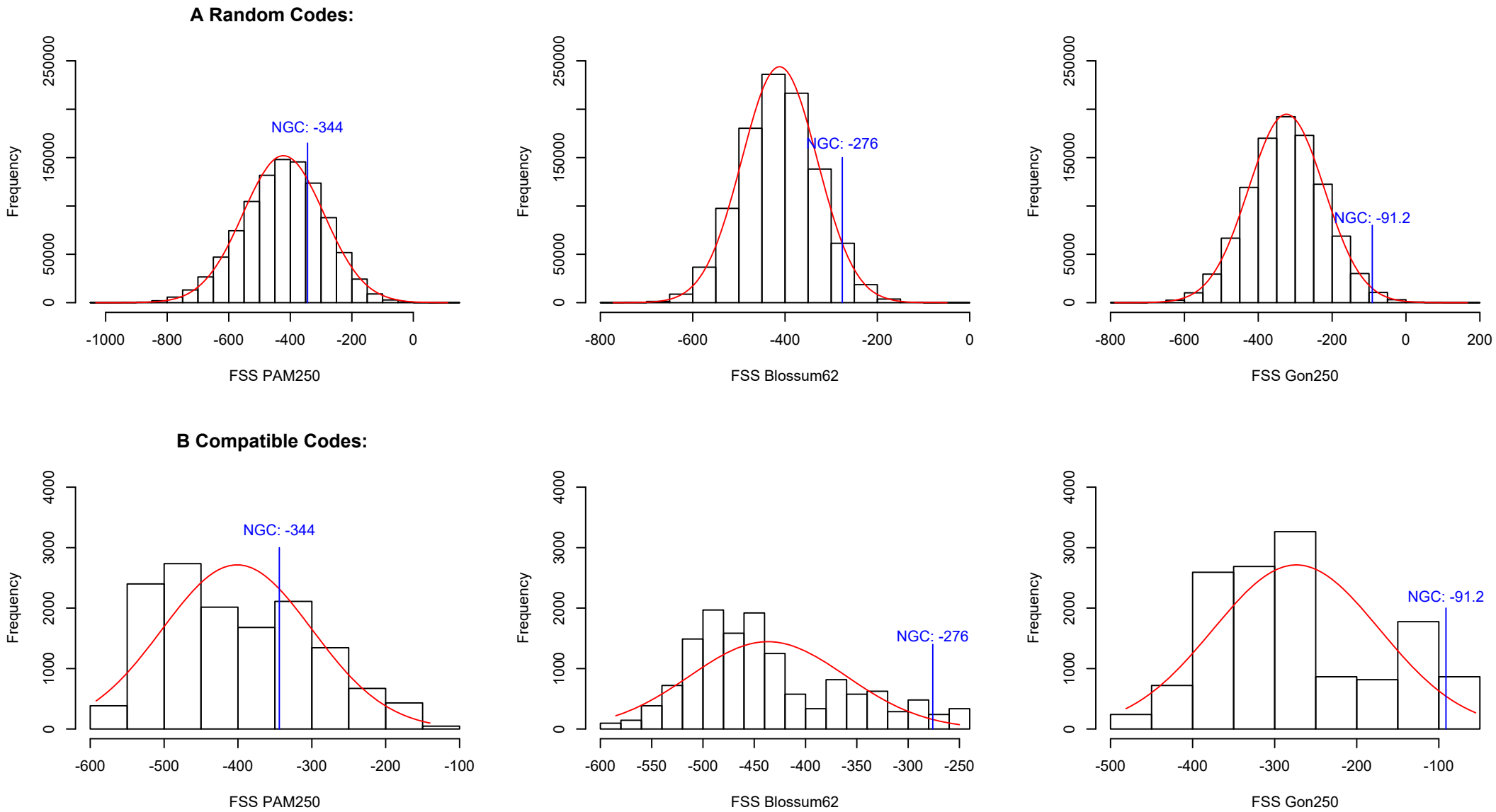
```
                        *        20         *
vegfaa     : MNLVVYLIQLFLAALLHLSAVKAAHIPKEGGKSKNDVI :  38
vegfaa-1   : MTWLFI*YSYFSRLSSICLL*RLPTYPKKGERAKMM*F :  35
vegfaa-1-r : MTWLFIWYSYFSRLSSICLLKRLPTYPKKGERAKMMWF :  38
vegfaa-2   : -MLGCLFDTVISRGSPPSVCCKGCPHTQRRGKEQK*CD :  36
vegfaa-2-r : -MLGCLFDTVISRGSPPSVCCKGCPHTQRRGKEQKWCD :  37
                         sr s      4         4

           40         *        60         *
vegfaa     : PFMDVYKKSACKTRELLVDIIQEYPDEIEHTYIPSCVV :  76
vegfaa-1   : PSWMCIKRVRARPESCW*TSSRSIPMRSSTRTSRPVWF :  72
vegfaa-1-r : PSWMCIKRVRARPESCWSTSSRSIPMRSSTRTSRPVWF :  76
vegfaa-2   : SLHGCV*KECVQDPRAAGRHHPGVSR*DRAHVHPVLCG :  72
vegfaa-2-r : SLHGCVKKECVQDPRAAGRHHPGVSRWDRAHVHPVLCG :  75
                c k4

             80         *        100         *
vegfaa     : LMRCAGCCNDEALECVPTETRNVTMEVLRVKQRVSQHN : 114
vegfaa-1   : SCAVQDAVMMRRSNASRQRHETSLWRCCGSSNAYRSII : 110
vegfaa-1-r : SCAVQDAVMMRRSNASRQRHETSLWRCCGSSNAYRSII : 114
vegfaa-2   : SHALCRML***GARMRPDRDTKRHYGGAAGQATRIAA* : 106
vegfaa-2-r : SHALCRMLKWWGARMRPDRDTKRHYGGAAGQATRIAAK : 113
             s a               r

               120         *        140         *
vegfaa     : FQLSFTEHTKCECRPKAEVKAKKENHCEPCSERRKRLY : 152
vegfaa-1   : FS*VSQNTPSVNAGQRQKSKQRKKTTVSLAQREGSACM : 147
vegfaa-1-r : FSWVSQNTPSVNAGQRQKSKQRKKTTVSLAQREGSACM : 152
vegfaa-2   : FSAEFHRTHQV*MQAKGRSQSKERKPL*ALLREKEALV : 142
vegfaa-2-r : FSAEFHRTHQVWMQAKGRSQSKERKPLWALLREKEALV : 151
             Fs     t  v    4  s  4         re  a

                  160         *        180
vegfaa     : VQDPLTCKCSCKFTQMQCKSRQLELNERTCRCEKPR- : 188
vegfaa-1   : CRTPSPVNAPANSHK-CNASPDNLS*TKELADVKSQD : 182
vegfaa-1-r : CRTPSPVNAPANSHK-CNASPDNLSKTKELADVKSQD : 188
vegfaa-2   : CAGPPHL*MLLQIHTNAMQVQTT*VKRKNLQM*KAKM : 176
vegfaa-2-r : CAGPPHLKMLLQIHTNAMQVQTTWVKRKNLQMWKAKM : 188
             c  P        h             4 l   K
```


A Random Codes:
NGC: -344
Frequency
FSS PAM250
NGC: -276
FSS Blossum62
NGC: -91.2
FSS Gon250
B Compatible Codes:
NGC: -344
FSS PAM250
NGC: -276
FSS Blossum62
NGC: -91.2
FSS Gon250