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

Xiaolong Wang[*1], *Quanjiang Dong*[2], *Gang Chen*[1], *Jianye Zhang*[1], *Yongqiang Liu*[1], *Yujia Cai*[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 had shown that frameshift genes/proteins are widely existing and sometimes functional. It is puzzling how a frameshift kept its structure and functionality while its amino-acid sequence is changed substantially. We revealed here that the protein sequences of the frameshifts are highly conservative when compared with the wild-type protein sequence, and the similarities among the three protein sequences encoded in the three 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 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., College of Life Sciences, 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]. 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 [2]; 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 reported that the standard genetic code was optimized for translational error minimization [3], which is being extremely efficient at minimizing the effects of mutation or mistranslation errors [4] and is optimal for kinetic energy conservation in polypeptide chains [5]. 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 [6].

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 [7, 8]. Therefore, frameshift mutations have been considered as mostly harmful and thus of little importance to the molecular evolution of proteins and their coding genes. However, frameshifted genes could sometimes be expressed through special mechanisms, such as translational readthrough, ribosomal frameshifting, or genetic recoding*, e.g.,* translational readthrough has been widely observed in various species [9, 10], and frameshifted coding genes can be corrected through programmed ribosomal frameshifting [11]. Moreover, frameshift homologs are widely observed [12], and frameshifted genes can be retained for millions of years and enable the acquisition of new functions [13], shed light on the role of frameshift mutations in molecular evolution.

Here, we report that frameshifts and wild-type protein sequences are always highly similar because the natural genetic code is optimal for frameshift tolerance. In addition, the genomes are also further optimized to tolerate frameshift mutations in certain species.

2. Results and Analysis

2.1 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 as meaningless, as they are not only totally different from the wild-type encoded in the main-frame but often interrupted by many stop signals. However, we noticed that frameshifted protein sequences and their wild-type counterparts are actually always highly similar [14], 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 [15]. As shown in Fig 1, a series of evolutionary events, including substitution, insertion, deletion, recombination, and several frameshifting events occurred in their *gp120* coding sequences (Fig 1A), and their encoded protein sequences are highly divergent but very similar (Fig 1B). These HIV/SIV strains were originated from a common ancestor [16], and their GP120 proteins are surely all functional, as the infection of these viruses into their host cells relies on these genes/proteins. In other words, the reading frames of *gp120* are changing in different strains, but their protein sequences are all highly similar and all functional. Hereafter, such frameshifts are called *functional frameshift homologs*.

As is well known, a protein can be dysfunctional even by changing one single residue, so, it is puzzling how a frameshift protein kept its tertiary structural and functional integrity while its primary sequence is being changed substantially.

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

To ask whether this phenomenon is unique or 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 counterpart are always highly similar in any given coding sequence. 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 physicochemical 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, which is arbitrarily selected and only for the purpose of visualization. In fact, one can easily reproduce the same kind of results by using any other coding sequences.

2.3 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 confirmed that δ is centered approximately at 0.5 in all CDSs (Table 2 and Supplementary Dataset 1). As shown in Table 2, the wild-type and the frameshift translations have a comparable amino-acid sequence similarity of 0.5 in all species tested, as well as in the simulated CDSs. In a word, the three protein sequences encoded in the three different reading frames are always similar, with an average pairwise 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.

2.4 The genetic code is optimal 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- and sequence-independent. This is also suggested by the simulated coding sequences, whose shiftability is comparable with those of the real coding genes. Therefore, we speculate that the shiftability is defined mainly by the genetic code rather than the coding sequences themselves.

As described in the method section, the averages of 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 the rest (94.5%) of them are all nonsynonymous. The proportion of positive nonsynonymous substitutions (29.7%), however, 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, the wobble codons are assigned mostly with synonymous substitutions, 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, *e.g.,* the average FSS (-1.78) is significantly higher than that of random codon substitutions (-10.81) (t-test P = 2.4969 × 10^{-10}), suggesting that 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 databases of protein sequences. 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 percentage of positive scores is only ~25%, *i.e.*, in random substitutions, the percent of positive substitutions is ~25%, therefore, randomly generated protein sequences would have a similarity of ~25% on average.

However, as shown in Table 2, frameshifts and their wild-type protein sequences have a similarity of ~50%, which is two folds of the average similarity of random sequences. The ~50% similarities among the frameshifts and their wild-type protein sequences could be well explained by combining the similarity derived from frameshift codon substitutions (~35%) with the similarity derived from random codon substitutions (~25%), minus their intersection (~10%). Therefore, it is suggested that the shiftability of protein-coding genes is predetermined by the natural genetic code.

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

To investigate the degree of optimization of frameshift tolerability of the genetic code, we adopted two strategies to generated alternative codon tables:

(1) *Random codon tables,* as defined by Freeland and Hurst [4], which is produced by swapping the amino acids assigned to sense codons while keeping all of the degenerative codons synonymous; The total number of all possible random codon tables is 20! = 2.43290201×10^{18}. Using their methods, we produced one million random codon tables, computed their FSSs (Supplementary Dataset 3), sorted and compared with those of the natural genetic code.

(2) *Alternative codon tables,* as created by Itzkovitz and Alon [6], which is produced by permuting the bases in the three codon positions independently and preserving the amino acid assignment. For each codon position, there are 4! = 24 possible permutations of the four nucleotides. So, there are 24^3 = 13,824 alternative codon tables.

Using their methods, we produced all 13,824 alternative codon tables, computed their FSSs (Supplementary Dataset 3), sorted and compared with those of the natural genetic code. As shown in Fig 3 and Table 5, the FSSs of the natural genetic code ranks in the top ~30% of random or compatible genetic codes when their FSSs are computed using scoring matrix PAM250, but ranks in the top 1.0–5.0% of random or compatible genetic codes when their FSSs are computed using BLOSSUM62 and GON250. It is known that the scoring matrices BLOSSUM and GON are newer and far more accurate than PAM, the oldest substitution scoring matrices. The results given by BLOSSUM62 and GON250 are also highly consistent with each other, therefore, we conclude that the FSS of the natural genetic code ranks in the top 1.0–5.0% of all possible alternative codon tables.

As described by Itzkovitz and Alon [6], by imposing the wobble constraint for base pairing in the third position, only two permutations are allowed in the third position: the identity permutation and the A↔G permutation. Therefore, the ensemble of alternative codes contains only 24 × 24 × 2 = 1152 distinct codes. In these 1152 alternative codes, only a dozen or dozens are better than the natural genetic code in terms of frameshift tolerability, the genetic code is therefore truly optimal in terms of frameshift tolerance.

2.6 The shiftability was 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 a shiftability value much higher than the average. In other words, although the shiftability of a certain gene is determined mainly by the genetic code, it could also be adjusted at the sequence level. We thought that biased usages of codons may contribute to sequence-level shiftability. As shown in Table 6 and Supplementary Dataset 4, the average FSS weighted by their codon usages are lower than the expected average FSS of the unbiased usage of codons in *E. coli, A. thaliana,* and *C. elegans*, showing that frameshift-tolerable codons are not preferred in their genomes. However, weighted average FSSs are significantly higher than the average FSS of the unbiased usage of codons in human, mouse, *Xenopus,* and yeast, suggesting that frameshift-tolerable codons are overrepresented in these species.

On the other hand, the usages of codon pairs are also highly biased [17], and the usage of codon pairs may also influence the shiftability of coding genes. As shown in Table 7 and Supplementary Dataset 5, the usages of codon pairs are biased in all tested species. Surprisingly, less than one-third (up to 1660) of the 4096 possible codon pairs are over-represented. The rest two-third (>2400) 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,* the weighted average FSS of their codon pairs usages are lower than the expected average FSSs of unbiased usage of codon pairs, showing that frameshift-tolerable codon pairs are not preferred in these genomes. In human, mouse, *Xenopus*, and yeast, however, their weighted average FSSs are significantly higher than the expected average FSSs of unbiased usage of codon pairs, showing that frameshift-tolerable codon pairs are overrepresented in these genomes.

3. Discussion

The natural genetic code has existed since the origin of life and was thought to have been optimizing through competition with other possible codes [18]. The natural genetic code was optimized along with several properties during the early history of evolution [19]. It was reported that the natural genetic code is optimal for translational error minimization, which is explained by the selection to minimize deleterious effects of translation errors [3]. In addition, it has been reported that only one in every million alternative genetic codes is more efficient than the natural code in terms of minimizing the effects of point-mutations or translational errors [4], and that the natural genetic code is nearly optimal for allowing additional information within coding sequences, such as out-of-frame hidden stop codons (HSCs) [6].

In this study, we discovered that the code-level shiftability of coding genes guaranteed on average half of the sites are kept conserved in a frameshift protein when compared with wild-type protein sequences. This underlying design of the natural genetic code forms a theoretical basis of frameshift tolerance, makes it understandable why functional frameshift homologs are widely observed [12]. Proteins have been and are evolving through point and frameshift mutations in their coding genes. 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 a more effective means to change protein sequences rapidly for the fast-evolving of novel or overlapping genes.

A complete frameshift is usually malfunctioned, and functional frameshifts are mostly partial frameshifts. Because of the shiftability of genes, in partial-frameshift coding genes, half of the frameshift codons and all of the non-frameshift codons are conservative. Thus, the protein sequences encoded by a partial frameshift is always even more similar to the wild-type than that by the complete frameshift, no matter where the partial frameshift starts and ends. There is no guarantee the proper functioning of any frameshifts, however, partial frameshifts can have the best chance to recover their structure and function in general. And therefore, the natural genetic code with the best shiftability could have the best chance to win the competition with the other genetic codes in the earlier evolution history.

There have been quite a few hypotheses on the causes and consequences of the usages of codons or codon pairs, such as gene expression level [20], mRNA structure [21], mRNA stability [22, 23], and protein abundance [24]. Here we demonstrated that gene-/genome-level shiftability is achieved through a biased usage of codons and dicodon, 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/dicodon. The over-represented frameshift-tolerable codons and codon pairs could possibly 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.

Here, we analyzed the shiftability of protein-coding genes in some model organisms. It will be interesting to investigate the shiftability of protein-coding genes in other species. Conceivably, the shiftability of protein-coding genes could possibly play an important role in the functioning, repairing and evolving of the proteins and their coding genes. Finally, the frameshifts are not assumed to be tolerated in the sense that they are not changed. If they are not removed by selecting against, they are assumed to be preserved or repaired in the evolutionary history [25].

4. Conclusion

Through the above analysis, we conclude that the natural genetic code is optimal in terms of shiftability (frameshift tolerability). The shiftability of coding genes guarantees a half-conservation of frameshifts, endows all and any proteins, coding genes and organisms inherent tolerability to frameshift mutations, owing to an adaptive advantage for the natural genetic code. As the “bottom design”, the genetic code allows coding genes and genomes for all species to tolerate frameshift mutations in both forward and backward directions, and thus, had 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 tolerate frameshift mutations occurred in protein-coding genes, by which the disasters of frameshift mutations have been utilized as a driving force for molecular evolution.

5. Materials and Methods

5.1 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 [26]. 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 [27] using default settings.

5.2 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$\geqslant$0) by using the scoring matrix GON250. In the alignments, each position of a gap or a negative score was counted as a difference.

5.3 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:

$$\boldsymbol{B}_1 \boldsymbol{B}_2 \boldsymbol{B}_3 \mid B_4 B_5 B_6 \mid \boldsymbol{B}_7 \boldsymbol{B}_8 \boldsymbol{B}_9 \mid \ldots \mid \boldsymbol{B}_{3i+1} \boldsymbol{B}_{3i+2} \boldsymbol{B}_{3i+3} \mid B_{3i+4} B_{3i+5} B_{3i+6} \mid \ldots \mid B_{3n-2} B_{3n-1} B_{3n}$$

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:* $\boldsymbol{B}_2 \boldsymbol{B}_3 B_4 \mid B_5 B_6 \boldsymbol{B}_7 \mid \ldots \mid \boldsymbol{B}_{3i+2} \boldsymbol{B}_{3i+3} B_{3i+4} \mid B_{3i+5} B_{3i+6} \boldsymbol{B}_{3i+7} \mid \ldots$

(2) *Delete two:* $\boldsymbol{B}_3 B_4 B_5 \mid B_6 \boldsymbol{B}_7 \boldsymbol{B}_8 \mid \ldots \mid \boldsymbol{B}_{3i+3} B_{3i+4} B_{3i+5} \mid B_{3i+6} \boldsymbol{B}_{3i+7} \boldsymbol{B}_{3i+8} \mid \ldots$

(3) *Insert one:* $B_0 B_1 B_2 \mid \boldsymbol{B}_3 B_4 B_5 \mid B_6 \boldsymbol{B}_7 \boldsymbol{B}_8 \mid \ldots \mid \boldsymbol{B}_{3i+3} B_{3i+4} B_{3i+5} \mid B_{3i+6} \boldsymbol{B}_{3i+7} \boldsymbol{B}_{3i+8} \mid \ldots$

(4) *Insert two:* $B_{-1} B_0 \boldsymbol{B}_1 \mid \boldsymbol{B}_2 \boldsymbol{B}_3 B_4 \mid B_5 B_6 \boldsymbol{B}_7 \mid \ldots \mid \boldsymbol{B}_{3i+2} \boldsymbol{B}_{3i+3} B_{3i+4} \mid B_{3i+5} B_{3i+6} \boldsymbol{B}_{3i+7} \mid \ldots$

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 [28], PAM250 [29-31], or GON250 [32]. 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.

5.4 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 [4], 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! = 2.43290201×10^{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 [6], 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.

5.5 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 [33]. 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.

5.6 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, $\boldsymbol{B_1 B_2 B_3} \mid B_4 B_5 B_6$, where, $B_k = \{A, G, U, C\}$, $k = 1 \ldots 6$. There are 400 different amino acid pairs and 4096 different codon pairs.

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.*

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

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

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.

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

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 [34, 35]. Translational readthrough could occur upon the activity of a suppressor tRNA with an anticodon matching a nonsense codon. Nonsense suppression tRNAs reported in *E. coli* includes *amber suppressors* (*supD* [36], *supE* [37], *supF* [38]), *ochre suppressor*s (*supG* [39]), and *opal suppressors* (*supU* [38], *su9*[40]).

In this study, these suppressor tRNAs were summarized as a set of *readthrough rules* (Table 1). These readthrough rules are used in the translating of artificial frameshift coding sequences. However, readthrough rules are not taken as *biological laws*, but *computational methods borrowed from biology*. The purpose of computational reading through is to obtain consecutive frameshift translations without the interruption of stop signals. The frameshift protein sequences translated from an artificial frameshift CDS do not really exist in biology but are used only for the alignment and compute the similarities of the three possible protein sequences encoded in the three different reading frames of the CDS. Therefore, the artificial frameshifting and reading-through operations we performed here 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 expected to be negligible, and this 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 average similarities of the frameshifts and the wild-type for readthrough and non-readthrough translations are 0.5354 and 0.5573, respectively.

Coding sequence data are available at GenBank, Ensembl or UCSC Genome Database. Code used to analyze the data can be found at https://github.com/CAUSA/Frameshift. Supplemental files available at FigShare (https://doi.org/10.6084/m9.figshare.9948050.v1). File S1 contains frameshift similarity data; File S2 contains frameshift substitutions scores of the natural genetic code; File S3 contains frameshift substitutions scores of the alternative genetic codes; File S4 contains frameshift substitutions scores of different usages of codons; File S5 contains frameshift substitutions scores of different usage of codon pairs.

Author Contributions

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

Acknowledgments

This study was funded by the *National Natural Science Foundation of China* through Grant 31571369. We acknowledge Dr. Hervé Seligmanna and Prof. Michael Lynch, for their comments on the preprints of this article [14, 41].

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 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 ClustalW2, show in GeneDoc with the amino acids 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 alternative codon tables. NGC: the natural genetic code; 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 or simulated proteins and their frameshift forms.

No.	***Species***	***Number of***	***Average Similarity***					
		CDSs	$\boldsymbol{\delta_{12}}$	$\boldsymbol{\delta_{13}}$	$\boldsymbol{\delta_{23}}$	$\boldsymbol{\delta}$	***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 weighted 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 weighted 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

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

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

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

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

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

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

12. 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.

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

14. 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.

15. 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**.

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

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

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

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

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

21. 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.

22. 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.

23. 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.

24. 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.

25. 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.

26. 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.

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

28. 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.

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

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

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

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

33. 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.

34. 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.

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

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

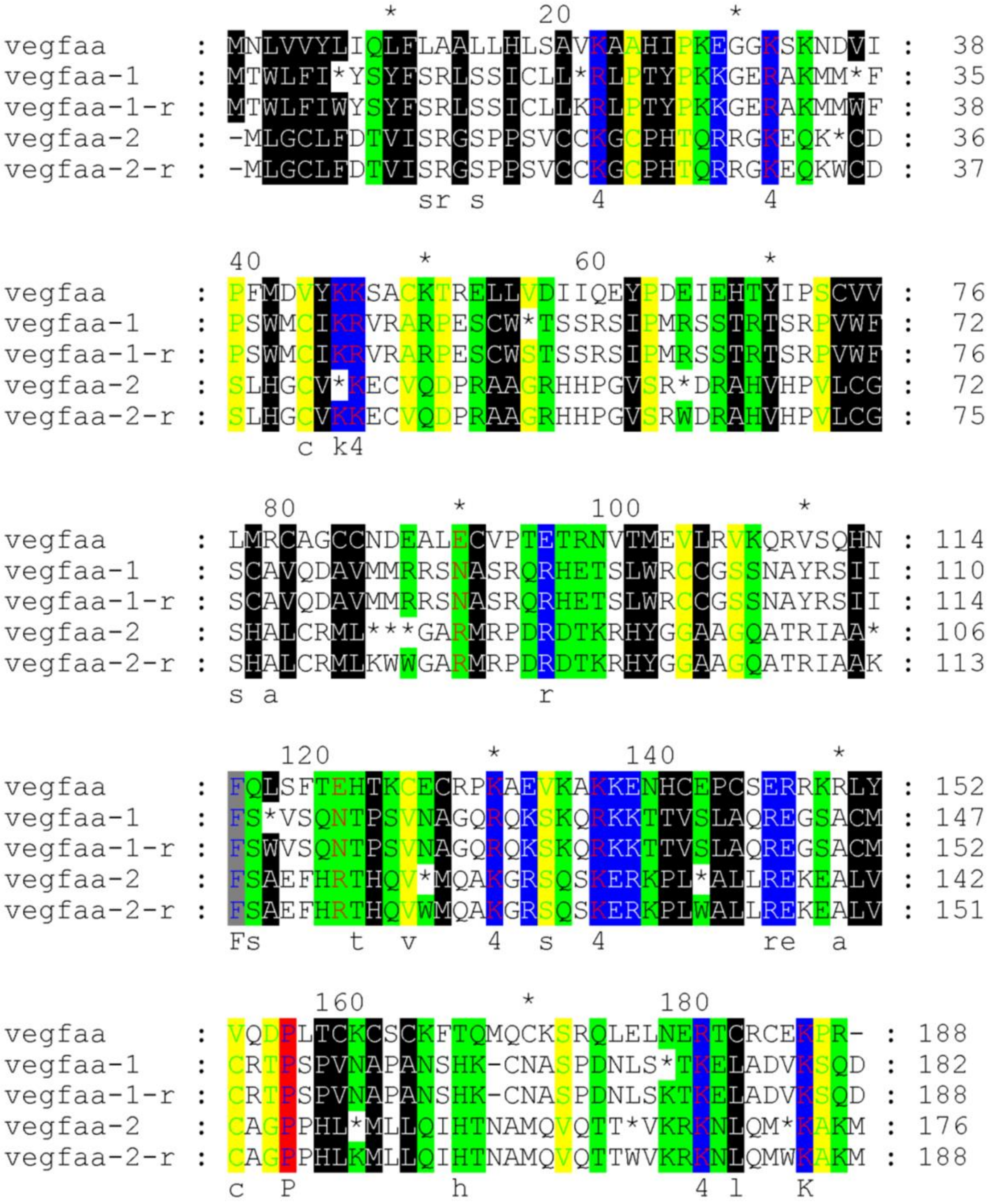
41. Wang, X., et al., *Why are frameshift homologs widespread within and across species?* bioRxiv, 2016.

```
                 *        20         *        40
HV1J3 : -----------ATGAGAGTGAAGGGGATCAGGAAGAA--TTA :   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
```

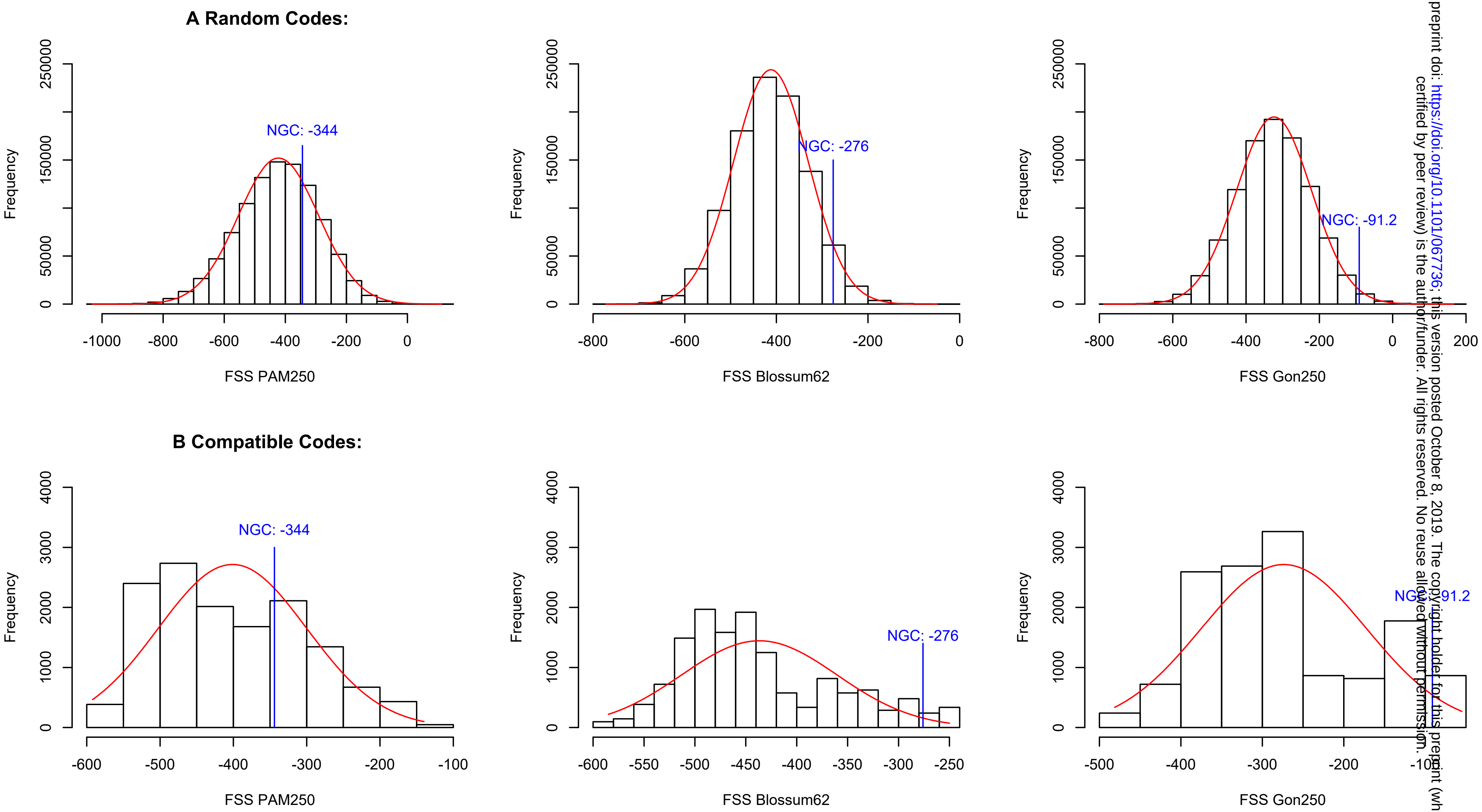
Fig 1 (A). Alignment of coding sequences of HIV/SIV GP120

```
                 *        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:
Frequency
0
50000
150000
250000
NGC: -344
-1000
-800
-600
-400
-200
0
FSS PAM250
NGC: -276
-800
-600
-400
-200
0
FSS Blossum62
NGC: -91.2
-800
-600
-400
-200
0
200
FSS Gon250
B Compatible Codes:
0
1000
2000
3000
4000
NGC: -344
-600
-500
-400
-300
-200
-100
FSS PAM250
NGC: -276
-600
-550
-500
-450
-400
-350
-300
-250
FSS Blossum62
NGC: -91.2
-500
-400
-300
-200
-100
FSS Gon250