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. Here we demonstrate 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 higher species, the shiftability is further optimized at gene- or genome-level by a biased usage of codons and codon pairs, in which frameshift-tolerable codons/codon pairs are overrepresented in their genomes.

[1] To whom correspondence should be addressed: *Xiaolong Wang, 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. Background

The genetic code was deciphered in the 1960s [1]. The standard genetic code consists of 64 triplet codons: 61 sense codons for 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/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 coding genes. However, frameshifted genes could sometimes be corrected through translational readthrough [9, 10], programmed ribosomal frameshifting [11], or genetic recoding [12]. Moreover, reading frame transitions are widely observed in prokaryotic genomes [13], and frameshifted genes can be retained for millions of years and enable the acquisition of new functions [14].

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, in certain species, their genome sequences are further optimized to tolerate frameshift mutations, shed light on the role of frameshift mutations in molecular evolution.

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). 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 CDSs, each containing 500 random sense codons, were simulated by *Recodon* 1.6.0 [15] using default settings.

2.2 Aligning and computing the similarity of the wildtype and frameshifts

Program Frameshift-Align.java was written to batch translate all coding sequences in three reading frames, align the 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 wildtype 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:* $\boldsymbol{B_3} B_4 B_5 \mid B_6 \boldsymbol{B_7} \boldsymbol{B_8} \mid \ldots \mid \boldsymbol{B_{3i+3}} B_{3i+4} B_{3i+5} \mid B_{3i+6} \boldsymbol{B_{3i+7}} \boldsymbol{B_{3i+8}} \mid \ldots$

(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 which were denoted as *FF* and *BF*. In either case, in most codons, 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 main types: (1) *Synonymous substitution* (SS), (2) *Nonsynonymous substitution* (NSS). Based on the above analysis, we further classified codon substitutions into three subtypes: (1) *Random substitution*: randomly change one, two, or three of the three bases of the codons, including 64 × 64 = 4096 possible codon substitutions; (2) *Wobble substitution*: randomly change only the third position of the codons, including 64 × 4 = 256 possible codon substitutions; (3) *Frameshift substitution*: substitutions caused by forward or backward frameshifting, 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). Program Frameshift-CODON.java, was written to compute the average substitution scores for each kind of codon substitution by using a scoring matrix, BLOSSUM62 [16], PAM250 [17, 18], or GON250 [19]. Each of the 64 codons has 4 forward and 4 backward frameshift substitutions. By summing all 64 × 8 = 512 FSSs together, we get the total FSS of the genetic code, which is considered as a measurement of the frameshift tolerability of the genetic code.

2.4 Computational analysis of alternative codon tables

Program RandomCodes.java was used 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 ($20! = 2.43290201 \times 10^{18}$) 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 used to produce all possible (13824) kinds of alternative codon tables, which is 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.

2.5 Analysis of codon pairs and their frameshift substitution scores

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.

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

A java program, *Frameshift-CODONPAIR*, was written to compute 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.

2.6 Computational analysis of the usage of codon and codon pairs

The usages of codons/codon pairs for the abovementioned genomes were analyzed using the method proposed in reference [20]. 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/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.

2.7 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 nonsense codon [21]. There have been many studies reported that translational readthrough functions in *E. coli*, yeast and eukaryotes (including human), while the readthrough rules may vary among different species [22]. In *E. coli*, nonsense suppression tRNAs reported includes amber suppressors (*supD* [23], *supE* [24], *supF* [25]), ochre suppressors (*supG* [26]), and opal suppressors (*supU* [25], *su9* [27]).

In this study, these suppressor tRNAs were summarized as a set of readthrough rules (Table 1). These *in-silicon* 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 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. The purpose is to evaluate the impact of readthrough on the similarity calculations. Therefore, the 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 does not require or imply that these *in-silicon* readthrough rules must function in *E. coli* or any other species.

3. Results and Analysis

3.1 An overview 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 wildtype 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 always highly similar, 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. As shown in Fig 1, several frameshifting events, as well as a series of evolutionary events, including substitution, insertion, deletion, recombination, occurred in their *gp120* coding sequences (Fig 1A), and their encoded protein sequences are divergent but highly similar (Fig 1B). These HIV/SIV strains were originated from a common ancestor [28], 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* kept changing in different strains, but their protein sequences are highly similar and all functional,

In fact, it has been noticed that such kind of functional frameshift homologs are not rare but widely existing [13]. However, they are not well-studies but largely ignored. Only a few cases of functional frameshift homologs were reported so far. For example, by collecting human coding exons bearing insertions or deletions compared with the chimpanzee genome and identification of homologs from other species, Hahn and Lee identified nine frameshift homologs between human and chimpanzee, and some of them seems to be functional in both species [29].

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

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

To answer this question, 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 their pairwise similarities. Surprisingly, in all coding sequences tested, almost all of the alignments produce only one or a few gaps among the three translations, suggesting that the three translations are always highly similar. In other words, for any given coding gene, the two frameshift protein sequences and the wild-type counterpart are always highly similar. For example,

in the alignment of wild-type zebrafish VEGFAA with their frameshifts (Fig 2), 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 unexpected, and 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 with any other coding genes by using a common sequence alignment tool.

To evaluate the impact of readthrough/non-readthrough translation method on the alignment of wildtype/frameshifts and their similarity calculations, readthrough and non-readthrough frameshift translations were aligned with the wildtype. For example, the alignments of wild-type and frameshift VEGFAA protein sequences are fully consistent in readthrough and non-readthrough translations, except for the stop signals presented in the non-readthrough translations. The average similarities of the frameshifts and wildtype for readthrough and non-readthrough translations are 0.5354 and 0.5573, respectively. The expected proportion of nonsense codons is 3/64 = 4.69%, therefore, in principle the difference of similarity calculations between readthrough and non-readthrough translations is negligible, which is consistent with our observations.

3.3 The definition of the shiftability of the protein-coding genes

For a given CDS, the three translations from the three different frames in the sense strand, let δ_{ij} be the similarity between a pair of protein sequences encoded in frame *i* and frame *j* (*i, j=1,2,3,* $i \neq j$, $\delta_{ij} = \delta_{ji}$); here, the average pairwise similarity (percent of synonymous and conserved sites) among the three protein sequences is 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 is centered approximately at 0.5 in all CDSs (Table 2 and Supplementary Dataset 1). As shown in Table 2, the wildtype 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 highly like each other, 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.

3.4 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, as most (192/256=75%) of the wobble 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 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, 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 aa substitutions assigned to frameshift substitutions are significantly more conservative than those to the random 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 twice the average similarity of random sequences. The ~50% similarities among the frameshifts and their 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%), suggesting that the shiftability of coding genes is predetermined by the natural genetic code.

3.5 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, which is produced by swapping the amino acids assigned to sense codons while keeping all 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, created by Itzkovitz and Alon, 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 newer scoring matrices BLOSSUM and GON are far more accurate than PAM, the oldest 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, and therefore, the ensemble of alternative codes contains only 24 × 24 × 2 = 1152 distinct genetic codes. In these 1152 alternative codes, only a dozen (or up to 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.

3.6 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 real genes do have an average value slightly higher than that of the simulated ones. In other words, although the shiftability of genes depend mainly on the genetic code, the shiftability of a certain gene could be slightly adjusted at the sequence level. We thought that a biased usage 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 genomes.

On the other hand, the usages of codon pairs are also highly biased [30], 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. Again, 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 higher than the expected average FSSs of unbiased usage of codon pairs significantly, showing that frameshift-tolerable codon pairs are overrepresented in these higher species.

4. Discussion

The natural genetic code has existed since the origin of life and was thought to have been optimizing through competition with other genetic codes [31]. The natural genetic code was optimized along with several properties during the early history of evolution [32]. It was reported that the natural genetic code is optimal for the minimization of translational errors, 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 found 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 intriguing property of the natural genetic code forms the theoretical basis of frameshift tolerance, explains why functional frameshift

homologs are widely observed [13]. Proteins have been and will continue to be 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, structure, and function of a protein at a very slow rate. However, frameshift + point mutations provide a far more effective means to change protein sequences rapidly for the fast evolving of novel or overlapping genes.

A 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 the non-frameshift codons are conservative. Comparing with a complete frameshift, partial frameshifts are even more similar to the wildtype, regardless where is the starts and ends of the frameshift. There is no guarantee the proper functioning of any frameshifts, but the underlying shiftability mechanism makes frameshifts can have the best chance to recover normal structure and function in the process of gene repair. It is conceivable that the 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 some hypotheses on the cause and consequence of the usages of codons/codon pairs, such as gene expression level [33], mRNA structure [34], mRNA stability [35, 36], and protein abundance [37]. Here we demonstrated that gene-/genome-level shiftability is achieved through a biased usage of codons and codon pairs, suggesting that genomes were probably further optimized for frameshift tolerance in certain higher organisms. This suggests that the shiftability of coding genes could either be a cause or a consequence of the usages of codons/codon pairs. Comparing with unbiased codon/codon pairs usages, overrepresented frameshift-tolerable codons/codon pairs may have additional advantages in evolution or survival rate: upon a frameshift mutation, a frameshift would be more similar to its wild-type counterpart, and therefore, have better chances to keep or recover the normal function.

Here, we analyzed the shiftability of genes and genomes in some model organisms. It will be interesting to investigate the shiftability of genes and genomes in other species. 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 [38, 39], wherein the shiftability of coding genes may play a vital role in the functioning, repairing, and evolving of proteins.

5. Conclusion

Through the above analysis, we conclude that the natural genetic code is optimal in terms of shiftability (frameshift tolerability). The shiftability of protein-coding genes guarantees a half-conservation between wildtype and frameshifts, endows any proteins, coding genes, and organisms an 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 in any species to tolerate frameshift mutations, in both forward and backward directions, and therefore, 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 any coding genes, by which frameshift mutations, by themselves disasters, have been utilized as a driving force for molecular evolution.

Data accessibility. Coding sequences are available at GenBank, Ensembl or UCSC genome sequence databases. Java source code used to analyze the data are available at GitHub (https://github.com/CAUSA/Frameshift). Supplementary data are available at FigShare (https://doi.org/10.6084/m9.figshare.9948050.v1), in which File S1 contains frameshift similarities; 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. The preprint version of this article is available from the bioRxiv preprint server: doi: https://doi.org/10.1101/067736.

Authors' 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.

Competing interests. We declare that the authors have no competing interests.

Funding. This study was funded by *National Natural Science Foundation of China* through Grant 31571369.

Acknowledgements. We acknowledge Dr. Hervé Seligmanna for comments on the preprints of this article, helped improve the manuscript.

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 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 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 5. 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 6. 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	Over-represented	Under-represented	All
1	*H. sapiens*	1573	2523	50	-1.52	-7.80	-3.06
2	*M. musculus*	1505	2591	190	-2.83	-7.13	-3.81
3	*X. tropicalis*	1660	2436	148	-3.12	-6.98	-3.80
4	*D. rerio*	1493	2603	148	-4.87	-6.09	-5.18
5	*D. melanogaster*	1418	2678	140	-5.33	-5.86	-5.02
6	*C. elegans*	1469	2627	164	-6.47	-5.26	-6.11
7	*A. thaliana*	1566	2530	15	-6.30	-5.35	-6.37
8	*S. cerevisiae*	1493	2603	159	-4.86	-6.14	-4.27
9	*E. coli*	1389	2707	197	-6.76	-5.11	-6.82
10	*Equal Usage*	0	0	0	N/A	N/A	-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.* J Mol Evol, 1998. **47**(3): p. 238-48.
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 Res, 2007. **17**(4): p. 405-12.
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 Res, 2014. **42**(14): p. 8928-38.
10. Jungreis, I., et al., *Evidence of abundant stop codon readthrough in Drosophila and other metazoa.* Genome Res, 2011. **21**(12): p. 2096-113.
11. Chen, J., et al., *Dynamic pathways of -1 translational frameshifting.* Nature, 2014. **512**(7514): p. 328-32.
12. Russell, R.D. and A.T. Beckenbach, *Recoding of translation in turtle mitochondrial genomes: programmed frameshift mutations and evidence of a modified genetic code.* J Mol Evol, 2008. **67**(6): p. 682-95.
13. 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.
14. Raes, J. and Y. Van de Peer, *Functional divergence of proteins through frameshift mutations.* Trends Genet, 2005. **21**(8): p. 428-31.
15. Arenas, M. and D. Posada, *Recodon: coalescent simulation of coding DNA sequences with recombination, migration and demography.* BMC Bioinformatics, 2007. **8**: p. 458.
16. 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.
17. Dayhoff, M.O., *Computer analysis of protein evolution.* Sci Am, 1969. **221**(1): p. 86-95.
18. Dayhoff, M.O., *The origin and evolution of protein superfamilies.* Fed Proc, 1976. **35**(10): p. 2132-8.
19. Schneider, A., G.M. Cannarozzi, and G.H. Gonnet, *Empirical codon substitution matrix.* BMC Bioinformatics, 2005. **6**: p. 134.

20. Gutman, G.A. and G.W. Hatfield, *Nonrandom utilization of codon pairs in Escherichia coli.* Proc Natl Acad Sci U S A, 1989. **86**(10): p. 3699-703.

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

22. 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.

23. Hoffman, E.P. and R.C. Wilhelm, *Genetic mapping and dominance of the amber suppressor, Su1 (supD), in Escherichia coli K-12.* J Bacteriol, 1970. **103**(1): p. 32-6.

24. Kuriki, Y., *Temperature-sensitive amber suppression of ompF'-'lacZ fused gene expression in a supE mutant of Escherichia coli K12.* FEMS Microbiol Lett, 1993. **107**(1): p. 71-6.

25. 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.

26. 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.

27. 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.

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

29. Hahn, Y. and B. Lee, *Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences.* Bioinformatics, 2005. **21 Suppl 1**: p. i186-94.

30. Tats, A., T. Tenson, and M. Remm, *Preferred and avoided codon pairs in three domains of life.* BMC Genomics, 2008. **9**: p. 463.

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

32. Knight, R.D. and L.F. Landweber, *The early evolution of the genetic code.* Cell, 2000. **101**(6): p. 569-72.

33. Paul, P., A.K. Malakar, and S. Chakraborty, *Codon usage and amino acid usage influence genes expression level.* Genetica, 2018. **146**(1): p. 53-63.

34. 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.

35. Carneiro, R.L., et al., *Codon stabilization coefficient as a metric to gain insights into mRNA stability and codon bias and their relationships with translation.* Nucleic Acids Res, 2019. **47**(5): p. 2216-2228.

36. 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.

37. 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.

38. 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.

39. Strauss, B.S., *Frameshift mutation, microsatellites and mismatch repair.* Mutat Res, 1999. **437**(3): p. 195-203.

```
                    *         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

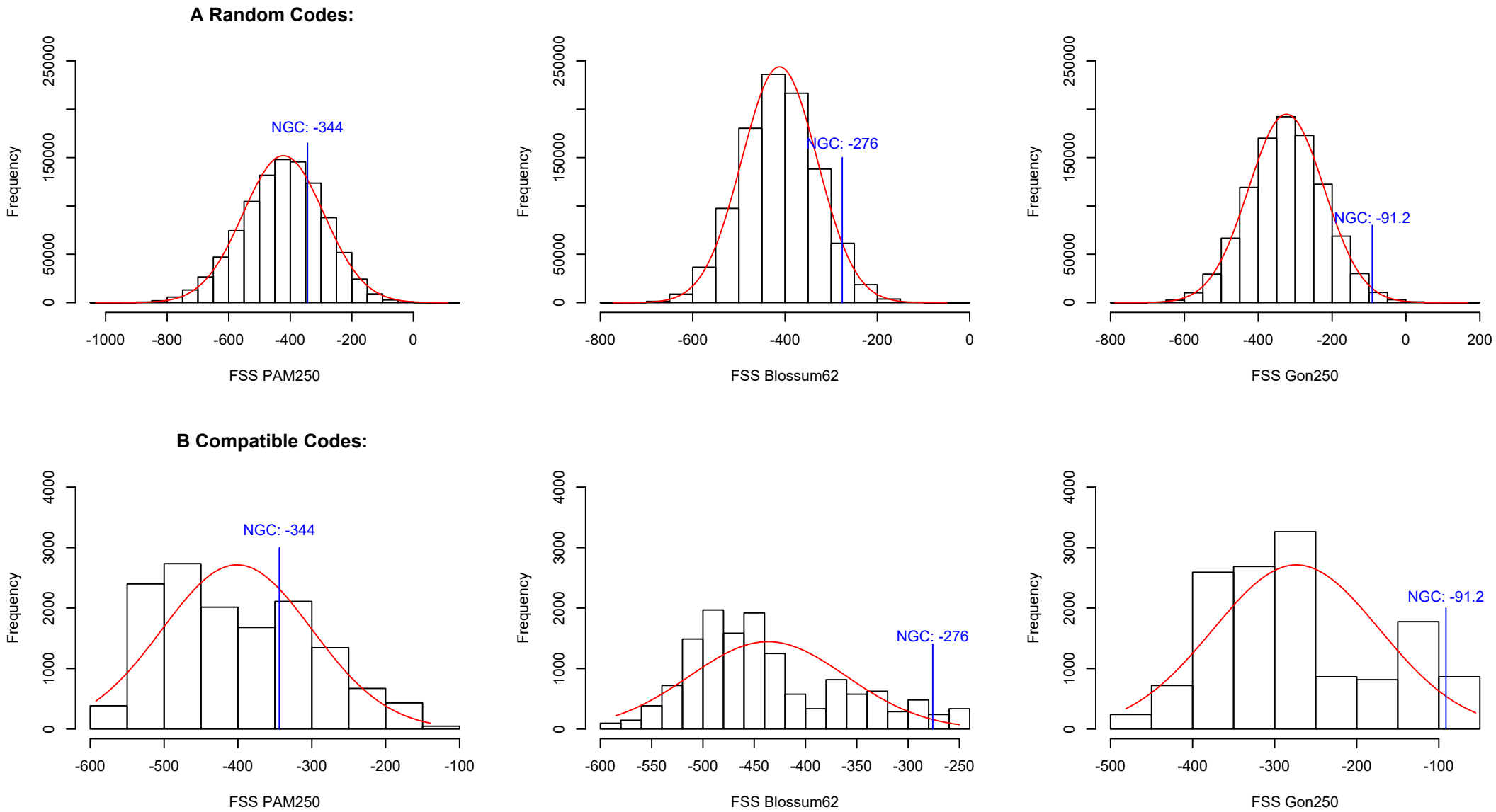
```
                        *        20         *
vegfaa     : MNLVVYLIQLFLAALLHLSAVKAAHIPKEGGKSKNDVI :  38
vegfaa-1   : MTWLFI*YSYFSRLSSICLL*RLPTYPKKGERAKMM*F :  35
vegfaa-1-r : MTWLFIWYSYFSRLSSICLLKRLPTYPKKGERAKMMWF :  38
vegfaa-2   : -MLGCLFDTVISRGSPPSVCCKGCPHTQRRGKEQK*CD :  36
vegfaa-2-r : -MLGCLFDTVISRGSPPSVCCKGCPHTQRRGKEQKWCD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