The universal genetic code, protein coding genes and genomes of all species were optimized for frameshift tolerance

Xiaolong Wang[*1], *Quanjiang Dong*[2], *Gang Chen*[1], *Jianye Zhang*[1], *Yongqiang Liu*[1], *Jinqiao Zhao*[1], *Haibo Peng*[1], *Yalei Wang*[1], *Yujia Cai*[1], *Xuxiang Wang*[1], *Chao Yang*[1], *Michael Lynch*[3]

1. *College of Life Sciences, Ocean University of China, Qingdao, 266003, P. R. China*
2. *Qingdao Municipal Hospital, Qingdao, Shandong, 266003, P. R. China*
3. *Center for Mechanisms of Evolution, BioDesign Institute, Arizona State University, Tempe, AZ, 85287-7701*

Abstract

Frameshifted coding genes yield truncated and dysfunctional peptides. Frameshift protein sequences encoded by the alternative reading frames of a coding gene have been considered as meaningless. And frameshift mutations have been considered as utterly harmful and of little or no importance for the molecular evolution of proteins. However, previous studies showed that frameshift coding genes can be expressed, and frameshift proteins can be functional by themselves. By analyzing all coding genes in nine model organisms, here we show that protein coding genes have a quasi-constant shiftability of 0.5: the frameshift protein sequences encoded in the alternative frames remain nearly half conservative when compared with the protein sequence encoded in the main frame. The shiftability of protein coding genes was predetermined mainly by the genetic code. In the universal genetic code, amino acid pairs assigned to frameshift substitutions are more conservative than those to random substitutions, and the frameshift tolerability of the standard genetic code ranks among the top 1.0-5.0% of all compatible genetic codes. In addition, in the genomes of all species tested, high frameshift-tolerance codon pairs are overrepresented, and thus, sequence-level shiftability are achieved by biased usages of codons and codon pairs. We concluded

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

that the genetic code, protein coding genes and genomes of all species were optimized to tolerate frameshift mutations.

1. Introduction

The genetic code was deciphered in the 1960s [1-4]. It 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: (1) The genetic code is universal for all species, with only a few variations found in some organelles or organisms, such as mitochondrion, archaea and yeast [5, 6]. (2) The triplet codons are redundant, degeneracy and wobble (the third base is interchangeable); (3) In an open reading frame, an insertion/deletion (InDel) causes a frameshift if the size of the InDel is not a multiple of three.

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

It was presumed that most frameshift protein-coding genes yield truncated, non-functional, potentially cytotoxic products, lead to waste of cell energy, resources and the activity of the biosynthetic machinery [11, 12]. Therefore, frameshift mutations were considered as harmful and of little importance to protein molecular evolution [13, 14]. Frameshift coding genes have been widely observed but generally considered as loss-of-function. However, it has been reported that frameshift genes are expressed through several special mechanisms, such as translational readthrough, ribosomal frameshifting and genetic recoding. For examples, it has been found that frameshift coding genes are tolerated in some animals by their translation systems in mitochondrial genes [15-17]; a (+1) frameshift insertion is tolerated in the *nad3* in birds and reptiles [15]. Moreover, frameshifted overlapping genes have been found in mitochondrial genes in fruit fly and turtles [18, 19]. In *E. coli*, high levels of

translational readthrough and ribosomal frameshifting have been characterized to be growth phase dependent [20]. Meanwhile, translational readthrough has also been widely observed in various species including yeast, fruit fly, fungi and mammals [21-27]. In addition, frameshift genes can be expressed through programmed translational/ribosomal frameshifting [28-31].

However, it has also been reported that some frameshift coding genes/proteins by themselves are functional in an alternative reading frame. For examples, (1) in yeast, a frameshift coding gene for mitochondrial cytochrome c oxidase subunit II (COXII), the sequence is translated in an alternative frame [32]; (2) in individuals having a frameshift mutation (138delT), alternative splicing led to the formation of a functional brain form variant enzyme of the pseudogene cytochrome that demethylates codeine to morphine [33]. Moreover, it was reported that frameshift mutations can be retained for millions of years and enable the acquisition of new gene functions [34]. In the last decade, a few studies have been reported that frameshift homologs are widely exist in many species [35], shed light into the role of frameshift mutation in molecular evolution.

As well known, proteins can be dysfunctioned even by changing only one residues, it is therefore a puzzle how these functional frameshift proteins kept their structures and functionalities while their amino acid sequences were changed substantially. Here we report that the standard genetic code, protein coding genes and genomes for all species were optimized at different levels to tolerate frameshift mutations.

2. Materials and Methods

2.1 Protein and coding DNA sequences

All available reference protein sequences and their coding DNA sequences (CDSs) in nine major 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* and/or *NCBI* Genome Databases. Ten thousand CDSs each containing 500

random sense codons were simulated by *Recodon* 1.6.0 using default settings [36]. 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 [37].

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

Program *Frameshift-Align*, written in java, was used to translate coding sequences in their three reading frames, align their three translations and compute their similarities. The standard genetic code was used to translate every CDS into three protein sequences in its three frames in the sense strand, but all internal nonsense codons were readthrough *in silicon* according to Table 1, the *in-vivo readthrough rules*. The wild-type and the two frameshift protein sequences were aligned by ClustalW2 using default parameters. Using the scoring matrix GON250, each position of a gap were counted as a difference, the pairwise similarity between each frameshift and its corresponding wild-type protein sequence is given by the percent of sites in which matched amino acids are conserved (amino acid substitution score ≥ 0).

2.3 Computational analysis of frameshift codon substitutions

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

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

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

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

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

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

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

So, 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*): $B_5 B_6 \boldsymbol{B_7}$ ($\rightarrow A_{22}$)

And so forth for each of the downstream codons. The results are two frameshifts, which were denoted as *FF* and *BF*. 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 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 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, each has 64×4=256 possible codon substitutions.

The amino acid substitution score of a frameshift codon substitution is defined as frameshift substitution score (FSS). A java program, *Frameshift-CODON*, was written to compute the average substitution scores for distinct kinds of codon substitutions by using a scoring matrix, BLOSSUM62 [38], PAM250 [39-41] or GON250 [42].

2.4 Computational analysis of random and alternative codon tables

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

Program *AlternativeCode.java* was used to produce 13824 compatible alternative codon tables proposed by Itzkovitz and Alon [10], by independently permuting the nucleotides in the three codon positions while preserving the amino acid assignment. Each alternative code has the same number of codons per each 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 genetic code were computed and sorted in ascending order, 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…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, the codon pair and its encoded amino acids has 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 usage of codons and codon pairs was analyzed on the above dataset using the same method used in reference [43]. The program *CODPAIR* was rewritten in java as the original program is unavailable. For each genome, it enumerates the total number of codons, and the number of occurrences for each codon and codon pair. The observed and expected frequencies were then calculated for each codon and codon pair. The result of these calculations is a list of 64 codons and 4096 codon pairs, each with an expected (E) and observed (O) number of occurrences, usage frequency, together with a value for $\chi_1^2 = (O - E)^2/E$. The codons and dicodons whose *O-value* is

greater/smaller than their *E-value* were identified as *over-/under-represented*, their average FSSs and the total weighted average FSSs were computed and compared.

3. Results and Analysis

3.1 The definition of frameshift homologs

A frameshift mutation often disrupts the function of a coding gene and its encoded protein, because every codon/aa is changed and often many stop codons emerge in the downstream. However, we noticed that the protein sequence encoded by a frameshifted CDS is often highly similar to the wild-type protein sequence. For example, different HIV/SIV strains, including HIV, SIVCZ and SIVGB, were originated from a common ancestor [44-46]. As shown in Fig 1A, the envelop glycoprotein coding gene (*gp120*) underwent a series of evolutionary events, including substitution, insertion, deletion, and recombination. Especially, several frameshifting events occurred in *gp120*, but their encoded GP120 protein sequences remain highly similar to each other (Fig 1B). These frameshift GP120 are surely all functional, as the infection of these virus into their host cells relies on these proteins.

As well known, a frameshift mutation is caused by one or more InDels in a protein coding gene whose length is not a multiple of three. Consequently, the reading frame is altered, either fully or partially. As abovementioned, frameshifted protein-coding genes have been widely observed, and some of them are actually functional by themselves. In this study, *frameshift homologs* are defined as a set of frameshifted but yet functional coding genes/proteins that were evolved from a common ancestor gene via frameshift mutation. Frameshift homologs are distinguishable from a pseudogene: a pseudogene usually contains a number of internal stop codons and is often dysfunctional or function through correction. Frameshift homologs, however, may or may not contain internal stop codons, and is a protein coding gene that are frameshifted but function normally.

3.2 Artificial frameshift protein sequences are always highly similar to the wild-types

The protein sequences encoded in the alternative reading frames of a coding gene (frameshifts) are generally considered as meaningless, as they are obviously different from the wild-type in the main-frame and are often interrupted by some stop signals. As mentioned above, however, we noticed that the frameshifts are highly similar to the wild-type if the stop signals were ignored. To validate whether or not this phenomenon is universal, all of the coding genes for nine model organisms were translated each into three protein sequences in their three different reading frames in the sense strand, and then each of the three translations were aligned by ClustalW. Their pairwise similarities were computed. Each position in the gap was counted as a difference. Incredibly, in all coding genes tested, the alignments of their three translations produce no or only a few gaps, showing that the three translations are highly similar to each other actually. For an 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 kept conserved in their physiochemical properties. Here we must point out that this example is nothing special but very common. It was not cherry picked but arbitrarily selected for visualization.

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 was centered approximately at 0.5 in all CDSs, in all species tested, as well as in the simulated CDSs (Table 2 and Supplementary Dataset 2). As shown in Table 2, in all species, most of their coding genes have a comparable shiftability of 0.5. In other words, in most coding genes, the three protein sequences encoded in their three different frames are always highly similar to each other, with an average similarity of ~50%. Therefore, we propose that *protein coding genes have a quasi-constant*

shiftability, approximately equals to 0.5. In other words, in any coding gene, on average nearly half of its amino acids remain conserved in the frameshifts, forming the basis of frameshift tolerance of the genetic code and the protein coding genes.

3.3 The readthrough rules and their impact on computation

The *in-vivo readthrough rules* (Table 1) were summarized from known nonsense suppression tRNAs reported in *E. coli*. The suppressor tRNAs are expressed to correct nonsense mutations, including *amber suppressors* (*supD* [47], *supE* [48], *supF* [49]), *ochre suppressor*s (*supG* [50]) and *opal suppressors* (*supU* [49], *su9* [51]). These suppressor tRNAs are taken as *in-silicon readthrough rules.*

Translational readthrough could occur upon activity of a suppressor tRNA with an anticodon matching a stop codon. The underlying causes for translational readthrough vary among species or studies. The suppressor tRNAs frequently occur in the negative strand of a regular tRNA [52-54]. It was found that translational readthrough occurred by using these suppressor tRNAs allows the translation of off-frame peptides [55-58]. There also have been many studies reported that translational readthrough functions in *E. coli*, yeast and many eukaryotes species (including human), while the readthrough rules may vary among different species [59, 60]. In addition, there have been increasing evidences showing that translational readthrough is linked to ribosomal frameshifting. For example, the interaction of eRF3 with RNase L leads to an increased readthrough efficiency at premature termination codons and +1 frameshift efficiency [61].

However, in this study, the readthrough rules are not 'biological laws' but purely 'computational methods borrowed from biology'. The purpose is to obtain consecutive frameshift protein sequences without the interruption of stop signals. Therefore, the artificial frameshifting and *in silicon readthrough* operations performed on the coding sequences are practically different from the *in vivo translational readthrough*. The frameshift amino acid sequences translated from the artificially frameshifted CDSs are not really exist in biology but used only as inputs to ClustalW for multiple sequence alignment (MSA). The purpose of MSA is only to compute the similarities of the protein sequences encoded in the three reading frames.

We first evaluated the impact of readthrough and non-readthrough on the alignment of wild-type and frameshift protein sequences and the computation of their similarity. The readthrough and non-readthrough frameshifts were aligned with the wild-type by ClustalW, respectively. For example, as shown in Fig 2, the alignments of wild-type VEGFAA and frameshifts are the same in readthrough and non-readthrough translations, except for the stop signals presented in the non-readthrough alignments. The shiftability of *vegfaa* computed from readthrough and non-readthrough alignments is 0.5354 and 0.5573, respectively. The average proportion of nonsense codons of the total number of codons is only 3/64=4.69%, so the difference of similarities/shiftability computed from readthrough and non-readthrough translations/alignments is usually negligible. So, we translated the frameshift coding sequences by readthrough. In other words, *in silicon* readthrough is simply a computational operation and does not require or imply that these *in-silicon* readthrough rules must function in *E. coli* or any other species.

3.4 The genetic code was optimized for frameshift tolerance

In Table 2, the shiftability of the protein coding genes is similar in all species and all genes, and their standard deviation is very small, suggesting that the shiftability is largely sequence-independent, implies that the shiftability is predetermined mainly by the genetic code rather than defined by the gene/protein sequences. Otherwise, they should vary greatly, since the gene/protein sequences by themselves vary greatly. This is also suggested by the coding sequences simulated by Recodon, whose shiftability is comparable with those of the real coding genes.

As described in the method section, the average amino acid substitution scores for random, wobble and forward/backward frameshift codon substitutions were computed respectively. As shown in Table 3 and Supplementary Dataset 3, in all 4096 random codon substitutions, only a small proportion (230/4096=5.6%) of them are synonymous, and the proportion of positive codon substitutions is 859/4096=20.1%. In addition, most (192/230=83%) of the synonymous substitutions are wobble, and most (192/256=75%) of the wobble substitutions are synonymous. Thus, the average substitution score of the wobble substitutions is the highest. For frameshift

substitutions, only a small proportion (28/512=5.5%) of them are synonymous (Table 4), and the other 95.9% of them are all nonsynonymous. However, the proportion of nonsynonymous substitutions (29.7%) is about 1.5-fold of that of the random substitutions (20.1%), and about 2-fold of that of the wobble substitutions (15.6%). In summary, in the standard genetic code, the wobble codons are assigned mostly to synonymous substitutions, while frameshift substitutions are assigned more frequently to positive nonsynonymous codons substitutions.

In addition, no matter which substitution scoring matrix (BLOSSUM62, PAM250 or GON250) was used for computation, the average FSSs of the frameshift substitutions are always significantly higher than that of the random substitutions. For GON250, *e.g.*, the average FSSs of frameshift substitutions (-1.781) is significantly higher than that of random substitutions (-10.81) (t-test $P = 2.4969 \times 10^{-10}$), suggesting that the amino acid pairs assigned to the frameshift codon substitutions are significantly more conservative than those to the random codon substitutions.

Substitution scoring matrices are widely used to determine the pairwise similarities of amino acid sequences, to score alignments between evolutionarily related protein sequences, to search protein sequence databases, and so on. In these scoring matrices, it is well known that positive scores represent synonymous or similar aa substitutions, while negative scores stand for dissimilar ones. In commonly used substitution scoring matrices, such as BLOSSUM62, PAM250 and GON250, most of the substitution scores are negative and the percent of positive scores is only ~30%. Therefore, in the random codon/aa substitutions, the percent of positive aa substitutions is ~30%. However, as shown in Table 2, most of the frameshift protein sequences encoded in the alternative reading frames of the coding sequences have a ~50% similarity to the wild-type protein sequences: combining the similarity derived from frameshift substitutions (~35%) with the similarity from random substitutions (~25%), minus their intersection (~10%), well explains the ~50% similarities observed among the frameshift and the wild-type protein sequences. Therefore, it is

suggested that the shiftability of protein coding genes is predetermined by the genetic code and is largely independent on the protein/coding sequences themselves.

3.5 The natural genetic code ranks top in all possible alternative codon tables

To further investigate the optimization of frameshift tolerance of the natural genetic code, we generated alternative codon tables, computed their FSSs and compared with that of the standard genetic code. There are two strategies to generate alternative codon tables: (1) *random codon tables*, developed by Freeland and Hurst [8], is to change the amino acids assigned to sense codons randomly and keeping all of the degenerative codons synonymous; (2) *compatible codon tables*, proposed by Itzkovitz and Alon [10], is to permute the nucleotides in the three codon positions independently and preserving the amino acid assignment, so that each codon table has the same number of codons per each amino acid (and the same impact of misread errors) as in the standard genetic code.

The number of all possible random codon tables is $20! = 2.43290201\times10^{18}$, but that of the compatible codon tables is only $(4!)^3=13824$. Using their methods, we randomly selected one million random codon tables, and generated all of the compatible codon tables, computed and sorted the FSSs of these alternative genetic codes (Supplementary Dataset 6), as show in Fig 3 and Table 5, the FSSs of the natural genetic code ranks in the top ~30% in random and compatible genetic codes when they were computed using scoring matrices PAM250, but ranks in the top 1.0-5.0% of the random and compatible genetic codes when computed using scoring matrices BLOSSUM62 and GON250. It is well known that PAM is inaccurate, as it is the oldest substitution scoring matrices, and the scoring matrices (BLOSSUM and GON) are more accurate. Because the results computed from BLOSSUM and GON are not only better but also more consistent, we concluded that the FSS of the standard genetic codes ranks in the top 1.0-5.0% of all possible alternative codon tables, clearly demonstrate that the standard genetic code is nearly optimal in terms of frameshift tolerance, and therefore, the shiftability of protein coding genes is indeed defined by the genetic code.

3.6 The genetic code is symmetric in frameshift tolerance

The genetic code shows the characteristics of symmetry in many aspects [62-64], and it evolved probably through progressive symmetry breaking [65-67]. Here in all CDSs both forward and backward frameshifts have comparable similarities compared with the wild-type (Table 2). In addition, in the natural genetic code both forward and backward frameshift substitutions have the same number of SSs/NSSs and equal FSSs (Table 3). These data suggested that the natural genetic code is also symmetric in terms of shiftability and frameshift tolerance. This could also explain why the codons in the natural genetic code are not tetrad but triplet: triplet codon can be easily kept symmetric for both forward and backward frameshifting, while for tetrad codons the situations for frameshifting will be much more complicated. The symmetric frameshift tolerance of the genetic code is very important. In the real biological system, an asymmetric frameshift tolerance would be meaningless, because frameshift mutations of coding genes may occur in both directions.

3.7 The shiftability of genes is further optimized at sequence and genome level

Although the shiftability of a coding sequence is defined mainly by the genetic code, shiftability may also exist at the sequence level. Functionally important coding genes, such as housekeeping genes, which are more conserved, may also have greater shiftability when compared with other genes. At first, we thought that a biased usage of codons may contribute to the sequence-level shiftability. However, as shown in Table 6 and Supplementary Dataset 4, it is somewhat surprising that in *E. coli* and *C. elegans* the average FSSs weighted by their codon usages are even lower than for unweighted calculations (equal usage of codons). In the other species tested, although the weighted average FSSs are higher than for unweighted analyses, the difference is not statistically significant in all species tested ($P>0.05$), suggesting that the usage of codons has little or no direct impact on the shiftability. However, the usage of codons may influence the shiftability indirectly, *e.g.*, by shaping the pattern of codon pairs.

Given a pair of amino acids, $A_1 A_2$, which A_1 and A_2 have m_1 and m_2 synonymous codons, say $B_1B_2B_3$ and $B_4B_5B_6$, respectively. Therefore, the dicodon, $B_1B_2B_3|B_4B_5B_6$, has $m_1 \times m_2$ possible combinations, called *synonymous codon pairs* (SCPs). It has been reported that the usages of codon pairs are also highly biased in various species, such

as bacteria, human and animals [43, 68-72]. As shown in Table 7, and Supplementary Dataset 5, in all species tested, the usages of codon pairs are highly biased. Surprisingly, in these genomes, only 700~1000 codon pairs are over-represented, which are less than a quarter of the total number of possible combinations of codon pairs (4096). In addition, the average FSSs of the over-represented codon pairs are all positive, while those of the under-represented codon pairs are all negative; in addition, the weighted average FSSs of equal usage of codon pairs is negative, while that of biased usages of codon pairs are positive in all species tested, suggesting that a strong selective pressure has been acting on the usage of these synonymous codon pairs, so that high frameshift-tolerance codon pairs are more preferred in these genomes. Therefore, sequence-level shiftability does exist, and was achieved through a biased usage of codons and codon pairs, suggesting that the protein coding genes, as well as the genomes (exomes), are also optimized for frameshift tolerance. There have been many studies on the causes and consequences of the usage of codons, such as gene expression level [73-79], mRNA structure [80-86], protein abundance [86-89], and mRNA/protein stability [90-92]. The above analysis suggested that the usages of codon pairs is either the cause or the consequence of the shiftability of the protein-coding genes.

4. Discussion

4.1 The genetic code was optimized for frameshift tolerance

The natural genetic code results from selection during early evolution, and it was optimized along several properties when compared with other possible genetic codes [93-104]. It was reported that the natural genetic code was optimized for translational error minimization, because the amino acids whose codons differed by a single base in the first and third positions were similar with respect to polarity and hydropathy, and the differences between amino acids were specified by the second position is explained by selection to minimize the deleterious effects of translation errors during the early evolution of the genetic code [7]. In addition, it was reported that only one in every million alternative genetic codes is more efficient than the natural genetic code

in terms of minimizing the effects of point mutation or translation errors [8]. It was demonstrated that the natural genetic code is also nearly optimal for allowing additional information within coding genes, such as out-of-frame hidden stop codons (HSCs) and secondary structure formation (self-hybridization) [10].

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

In this study, we discovered that the code- and sequence-level shiftability of coding genes guaranteed on average over half of the sites are kept conserved in a frameshift protein when compared with the main protein sequences. This is the basis for frameshift tolerance, and an underlying design of the natural genetic code, explains why frameshift homologs are widely observed in many species. For partially frameshift coding genes, conservation is inversely proportional to the numbers of frameshift sites, therefore, partial frameshifts are all highly similar to their wild type. Hence, it is guaranteed that on average over half of their aa sites are kept conserved in the frameshift proteins when compared to their wild type. However, this does not mean that these frameshift variants are all functional, but some of them may maintain their structure or function. In addition, the wild type of a coding gene is not necessarily the best form but could have been changing through point or frameshift mutations. Point mutations improve or alter the structure and function of a protein at a very slow rate. However, frameshift + point mutations may provide faster and more effective means of molecular evolution for the generating of novel genes or the developing of overlapping genes. The frameshifts are not assumed to be tolerated in the sense that they are not degraded, or as functional as wild-type proteins, but they are not lost completely by selecting against, but could be preserved in the

evolutionary history, because they are assumed to be repairable [116-120]. We are further investigating how a frameshift mutation is repaired [121].

4.2 The universality of the shiftability

Here we analyzed the shiftability of protein-coding genes in some model organisms, thus it is interesting to validate this mechanism in other species. It has been reported that frameshift mutations are tolerated in some animals by their translation systems in mitochondrial genes [15-17]. For example, a (+1) frameshift insertion is tolerated in the *nad3* in birds and reptiles [15]. Moreover, frameshift overlapping genes have been found in mitochondrial genes in fruit fly and turtles [18, 19]. In *E. coli*, high levels of translational readthrough and ribosomal frameshifting have been characterized to be growth phase dependent [20]. Meanwhile, translational readthrough has been widely observed in various species including yeast, fruit fly, fungi and mammals [21-27]. And it has also been observed that frameshift genes can be expressed by programmed translational/ribosomal frameshifting [28-31]. In the last decade, a few studies have been reported that frameshift homologs are widely exist in many species [34, 35]. Conceivably, the shiftability of protein coding genes may contribute, at least partially, to the function, repair and evolution of frameshift proteins and their coding genes.

5. Conclusion

The natural genetic code have existed since the origin of life and may have been optimizing during early evolution through competition with other possible genetic codes [122-125]. Through the above analysis, we conclude that the natural genetic code is nearly optimal in terms of frameshift tolerance. Frameshift tolerance, translational readthrough and ribosomal frameshifting are widely observed in many species. As the “bottom design” for genes and genomes of all species, the universal genetic code allows all coding genes to tolerate frameshifting in both forward and backward directions, and thus has a better fitness in the early evolution. The shiftability of protein-coding genes guarantees a half-conservation of frameshift proteins, endows all coding genes and all organisms an inherent tolerability to

frameshift mutations, and suggests an adaptive advantage for the standard genetic code as it does eliminate the risk of drastic errors. Thanks to this ingenious property of the genetic code, the shiftability of coding genes serves as an innate mechanism for cells to deal with frameshift mutations, by which the disastrous frameshift mutations were utilized as a driving force for molecular evolution.

Author Contributions

Xiaolong Wang conceived the study, coded the programs, analyzed the data, prepared the figures, tables and wrote the paper; Yujia Cai, Jing Liu, Jing Lin, Yongqiang Liu and Jinqiao Zhao analyzed some data. Quanjiang Dong, Gang Chen and Jianye Zhang gave suggestions on the paper; Michael Lynch supported this study and revised the manuscript.

Acknowledgements

This study was funded by *National Natural Science Foundation of China* through Grant 31571369. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We acknowledge Dr. Hervé Seligmanna from Aix-Marseille University for his valuable comments and suggestions on the earlier preprints of this article [126, 127].

Additional Information

We declare that the authors have no competing interests.

Figure Legends

Fig 1. The alignment of the coding and the protein sequences of HIV/SIV GP120. (A) The alignment of GP120 coding sequences, with highlights showing that the coding genes contain several frameshifting events. In other words, the coding gene *gp120* is expressed in different reading frames in different virus strains. (B) The alignment of GP120 protein sequences, showing that the GP120 sequences for different virus, which are encoded in different reading frames of *gp120*, are highly similar. The alignment was aligned by ClustalW and show in GeneDoc with the bases/AAs colored by on their physicochemical property.

Fig 2. The ClustalW alignment of the wild-type VEGFAA and its readthrough or non-readthrough frameshifts.

Fig 3. The histogram of the FSSs for the genetic codes. (A) randomly chosen 1,000,000 random codon tables; (B) all 13824 compatible codon tables. FSSs were computed using scoring matrices PAM250, BLOSSUM62 and GON250, respectively. The probability densities were computed using the normal distribution function, $N(\mu, \sigma^2)$. The lines are plotted in language R.

1 Table 1. The natural suppressor tRNAs (*readthrough rules*) 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

2

3

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

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

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

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

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

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

Table 4. The synonymous frameshift substitutions

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

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

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

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

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

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

NO	*Species* (*Codon Usage*)	*Number of over-represented Codon pairs*	*Average FSS of over-represented Codon pairs*	*Average FSS of under-represented Codon pairs*	*Weighted Average FSS of All Codon pairs*
1	*H. sapiens*	712	41.30	-25.94	102.41
2	*M. musculus*	722	41.09	-26.09	98.55
3	*X. tropicalis*	725	42.20	-25.81	98.24
4	*D. rerio*	728	40.91	-26.17	87.38
5	*D. melanogaster*	723	39.77	-25.95	79.51
6	*C. elegans*	729	40.85	-26.18	81.48
7	*A. thaliana*	729	40.54	-26.09	90.64
8	*S. cerevisiae*	729	40.85	-26.18	99.21
9	*E. coli*	965	39.27	-30.75	77.03
10	*Equal Usage*	0	N/A	N/A	-28.50

2

3

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

35. 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.

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

37. 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.

38. 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.

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

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

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

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

43. 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.

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

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

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

47. 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.

48. 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.

49. 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.

50. 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.

51. 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.

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

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

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

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

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

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

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

59. 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.

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

61. Le Roy, F., et al., *A newly discovered function for RNase L in regulating translation termination.* Nat Struct Mol Biol, 2005. **12**(6): p. 505-12.

62. Findley, G.L., A.M. Findley, and S.P. McGlynn, *Symmetry characteristics of the genetic code.* Proc Natl Acad Sci U S A, 1982. **79**(22): p. 7061-5.

63. Frappat, L., P. Sorba, and A. Sciarrino, *Symmetry and codon usage correlations in the genetic code.* Physics Letters A, 1999. **259**(5): p. 339-348.

64. Koch, A.J. and J. Lehmann, *About a symmetry of the genetic code.* Journal of Theoretical Biology, 1997. **189**(2): p. 171-174.

65. Lenstra, R., *Evolution of the genetic code through progressive symmetry breaking.* J Theor Biol, 2014. **347**: p. 95-108.

66. Hornos, J.E.M., Y.M.M. Hornos, and M. Forger, *Symmetry and symmetry breaking: An algebraic approach to the genetic code.* International Journal of Modern Physics B, 1999. **13**(23): p. 2795-2885.

67. Antoneli, F. and M. Forger, *Symmetry breaking in the genetic code: Finite groups.* Mathematical and Computer Modelling, 2011. **53**(7-8): p. 1469-1488.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

86. 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.

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

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

89. 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.

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

91. 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.

92. 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

126. 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.

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

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

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

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

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

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

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

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

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

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

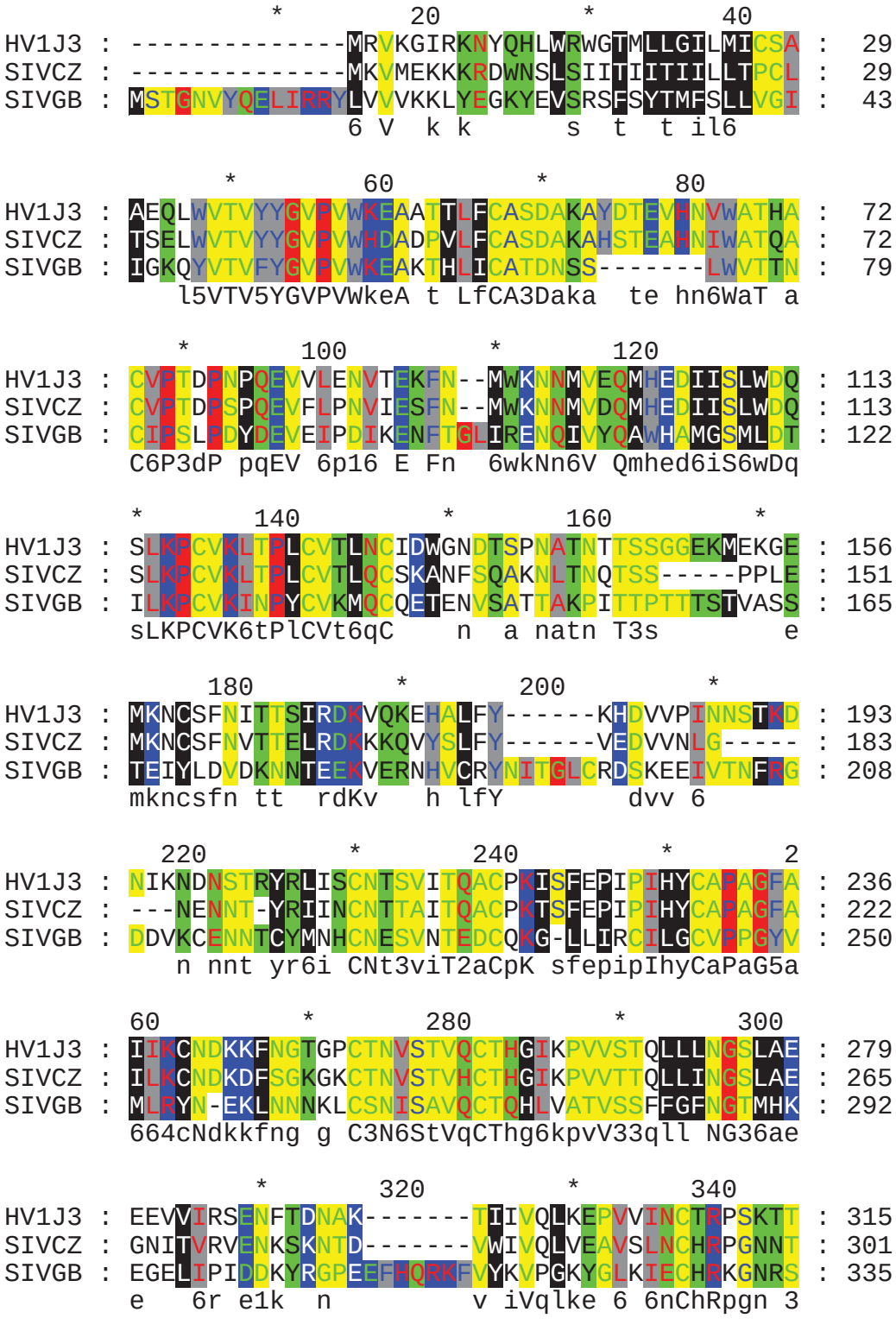

Fig 1 (B). Alignment of protein sequences of HIV/SIV GP120

Fig 2 alignment of VEGFAA and its frameshifts