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

Xiaolong Wang, Gang Chen, Jianye Zhang, Yongqiang Liu, Yujia Cai*

College of Life Sciences, Ocean University of China, Qingdao, 266003, P. R. China

Abstract

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

1. Introduction

The genetic code was deciphered in the 1960s [1]. The standard genetic code consists of 64 triplet codons: 61 sense codons for the twenty amino acids and the remaining three

nonsense codons for stop signals. The natural genetic code has several important properties: first, the genetic code is universal for all species, with only a few variations found in some organelles or organisms, such as mitochondrion, archaea, yeast, and ciliates [2]; second, the triplet codons are redundant, degenerate, and the third base is wobble (interchangeable); third, in an open reading frame, an insertion/deletion (InDel) causes a frameshift if the size of the InDel is not a multiple of three.

It has been revealed 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 harmful and of little importance to the molecular evolution of proteins and their coding genes. However, frameshifted genes could sometimes be expressed through special mechanisms, such as translational readthrough, ribosomal frameshifting, or genetic recoding*, e.g.,* translational readthrough has been widely observed in various species [9, 10], and frameshifted coding genes can be corrected through programmed ribosomal frameshifting [11]. Moreover, frameshift homologs are widely observed [12], and frameshifted genes can be retained for

millions of years and enable the acquisition of new functions [13], shed light on the role of frameshift mutations in molecular evolution.

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

2. Materials and Methods

2.1 Protein and coding DNA sequences

The human/simian immunodeficiency virus (HIV/SIV) strains were derived from the seed alignment in Pfam (pf00516). The coding DNA sequences (CDSs) of their envelop glycoprotein (GP120) were retrieved from GenBank. All available reference CDSs in nine model organisms were retrieved from UCSC Genome Database. Ten thousand simulated CDSs (each with 500 sense codons) were generated by *Recodon* 1.6.0 [14]. The details are described in the supplementary text (1.1).

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

Program *Frameshift-Align*.java was written to batch translate all coding sequences in their three different reading frames, align the three translations, and compute their pairwise similarities. Each internal nonsense codon was translated into an amino acid according to the *readthrough rules* (Table 1). The details are described in the supplementary text (1.2).

2.3 Computational analysis of frameshift codon substitutions

Program *Frameshift-CODON*.java is used to compute the average substitution scores for each kind of codon substitution by using a scoring matrix, BLOSSUM62, PAM250, or GON250. The details are described in the supplementary text (1.3).

2.4 Computational analysis of alternative codon tables

Program *RandomCodes.java* and *AlternativeCodes.java* were used to produce random codon tables [4] and compatible codon tables [6]. The sum of FSSs for each of the random or compatible genetic codes was computed and sorted ascendingly and compared with that of the natural genetic code. The details are described in the supplementary text (1.4).

2.5 Computational analysis of the usage of codon and codon pairs

The usages of codons and codon pairs for the abovementioned genomes were analyzed using the method proposed in reference [15]. The codons/dicodons whose observed value is greater/smaller than their expected value were identified as over-/under-represented, and their weighted average FSSs were calculated for each genome. The details are described in the supplementary text (1.5).

2.6 Analysis of the frameshift substitution scores of codon pairs

Program *CODONPAIR*.java was used 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, which is used to evaluate the frameshift tolerability of the codon pairs presented in a genome. The details are described in the supplementary text (1.6).

3. Results and Analysis

3.1 An example of functional frameshift homologs

As is known, a frameshift mutation is caused by one or more InDels in a protein-coding gene whose length is not a multiple of three, and thus the reading frame is altered, fully or partially. The frameshift protein sequences encoded in the alternative reading frames of a coding gene are generally considered as meaningless, as they are not only totally different from the wild-type encoded in the main-frame but often interrupted by many stop signals. However, we noticed that frameshifted protein sequences and their wild-type counterparts are actually always highly similar [16], 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 [17]. As shown in Fig 1, a series of evolutionary events, including substitution, insertion, deletion, recombination, and several frameshifting events occurred in their *gp120* coding sequences (Fig 1A), and their encoded protein sequences are highly divergent but still similar (Fig 1B). These HIV/SIV strains were originated from a common ancestor [18], and their GP120 proteins are surely all functional, as the infection of these viruses into their host cells rely on these genes/proteins. In other words, the reading frames of *gp120* are changing in different strains, but their protein sequences are very similar and all functional. Here, these frameshifts of GP120 are called *functional frameshift homologs*.

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

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

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

3.3 The definition of the shiftability of the protein coding genes

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

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

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

3.4 The genetic code is nearly 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/sequence independent. This is also suggested by the simulated coding sequences, whose shiftability is comparable with those of the real coding genes. We therefore speculate that the shiftability is defined mainly by the genetic code rather than the coding sequences themselves.

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

is 859/4096=20.1%. In addition, the average score of the wobble substitutions is the highest, because most (192/256=75%) of the wobble codon substitutions are synonymous, and most (192/230=83%) of the synonymous substitutions are wobble. For frameshift substitutions, only a small proportion (28/512=5.5%) of them are synonymous (Table S1) and all the other 94.5% of them are nonsynonymous. However, the proportion of positive nonsynonymous substitutions (29.7%) is ~1.5-fold of that of the random substitutions (20.1%), and ~2-fold of that of the wobble substitutions (15.6%). *In short, in the natural (standard) genetic code, wobble codons are assigned mostly with synonymous substitutions, while frameshift codon substitutions are assigned more frequently with positive nonsynonymous substitutions.*

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

Scoring matrices are widely used to determine the similarities of amino acid sequences, to score the alignments of evolutionarily related protein sequences, and to search databases of protein sequence. In all these scoring matrices, positive scores represent synonymous or similar aa substitutions, while negative scores stand for dissimilar ones. In commonly used scoring matrices, such as BLOSSUM62, PAM250, and GON250, most of the substitution

scores are negative and the average percent of positive scores is only ~25%, *i.e.*, in random substitutions, the percent of positive substitutions is ~25%, therefore, randomly generated protein sequences would have a similarity of ~25% on average.

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

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,* proposed by Freeland and Hurst [4], which is generated by swapping the amino acids assigned to sense codons while keeping all of the degenerative codons synonymous; (2) *compatible codon tables,* created by Itzkovitz and Alon [6], which is produced by permuting the bases in the three codon positions independently and preserving the amino acid assignment, so that each codon table has the same number of codons per each amino acid (and the same impact of misread errors) as in the standard genetic code. The total number of all possible random codon tables is an astronomical number, $20! = 2.43290201\times10^{18}$, while the total number of all compatible codon tables is $(4!)^3=13824$.

Using their methods, we produced one million random codon tables and all compatible codon tables, computed their FSSs (Supplementary Dataset 3), sorted and compared with those of the natural genetic code. As shown in Fig 3 and Table 4, the FSSs of the natural genetic code ranks in the top ~30% of random or compatible genetic codes when their FSSs are computed using scoring matrix PAM250, but ranks in the top 1.0–5.0% of random or compatible genetic codes when their FSSs are computed using BLOSSUM62 and GON250. It is known that the scoring matrices BLOSSUM and GON are newer and far more accurate than PAM, the oldest substitution scoring matrices. The results given by BLOSSUM62 and GON250 are also highly consistent with each other, therefore, we conclude that the FSS of the natural genetic code ranks in the top 1.0–5.0% of all possible alternative codon tables, suggesting that the genetic code is truly nearly 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 many genes do have a shiftability value much higher than the average. In other words, although the shiftability of a certain gene is determined mainly by the genetic code, it could be adjusted at sequence level. We thought that biased usages of codons may contribute to sequence-level shiftability. As shown in Table 5 and Supplementary Dataset 4, in *E. coli, A. thaliana* and *C. elegans,* the average FSSs weighted by their codon usages are lower than the expected average FSS of the unbiased usage of codons, showing that frameshift-tolerable codons are not preferred in their genomes. However, in human, mouse, *Xenopus,* and yeast, weighted average FSSs

are significantly higher than the average FSS of the unbiased usage of codons, suggesting that frameshift-tolerable codons are overrepresented in these species.

On the other hand, it has been reported that the usages of codon pairs are also highly biased [19]. The usage of codon pairs may also influence the shiftability of coding genes. As shown in Table 6 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 other two third (>2400) are under-represented or not even used, suggesting a strong selection acting on the usages of the synonymous codon pairs.

In *E. coli, C. elegans* and *A. thaliana,* the weighted average FSS of their codon pairs usages are lower than the expected average FSSs of unbiased usage of codon pairs, showing that frameshift-tolerable codon pairs are not preferred in these genomes. In human, mouse, *Xenopus*, and yeast, however, their weighted average FSSs are significantly higher than the expected average FSSs of unbiased usage of codon pairs, showing that frameshift-tolerable codon pairs are overrepresented in these genomes.

4. Discussion

The natural genetic code has existed since the origin of life and have been optimizing through competition with other possible codes during early history of evolution [20]. The natural genetic code was optimized along several properties [21]. It was reported that the natural genetic code is optimal for translational error minimization, which is explained by the selection to minimize deleterious effects of translation errors [3]. In addition, it has been reported that only one in every million alternative genetic codes is more efficient than

the natural code in terms of minimizing the effects of point-mutations or translational errors [4], and that the natural genetic code is nearly optimal for allowing additional information within coding sequences, such as out-of-frame hidden stop codons (HSCs) [6].

In this study, we discovered that the code-level shiftability of coding genes guaranteed on average half of the sites are kept conserved in a frameshift protein when compared with wild-type protein sequences. This underlying design of the natural genetic code forms a theoretical basis of frameshift tolerance, and explains why functional frameshift homologs are widely observed [12]. Proteins have been and are evolving through point and frameshift mutations in their coding genes. The rate of point mutation is extremely low, and so, point mutations alter the sequence, the structure, and the function of a protein at a very slow rate. However, (frameshift + point) mutations provide an effective means to change a protein sequence rapidly for the fast evolving of a novel or overlapping gene.

A large-size/complete frameshift is usually malfunctioned, and functional frameshifts are mostly small-size partial frameshifts. Because of the shiftability of genes, in partial-frameshift coding genes, conservative sites include half of the frameshift codons and all of the non-frameshift codons. Thus the protein sequence encoded by a partial frameshift is always more similar to the wild-type than that by the complete frameshift, no matter where the partial frameshift starts and ends. This does not guarantee the proper functioning of any partial frameshifts; however, partial frameshifts can have a greater chance to recover their structure and function in general.

There have been quite a few hypotheses on the causes and consequences of the usages of codons or codon pairs, such as gene expression level [22], mRNA structure [23], mRNA stability [24, 25], and protein abundance [26]. Here we demonstrated that gene-/genome-level shiftability is achieved through a biased usage of codons and dicodon, suggesting that genomes are further optimized for frameshift tolerance in certain organisms. This suggests that the shiftability of the protein-coding genes could be either a cause or a consequence of the usages of codons/dicodon. The over-represented frameshift-tolerable codons and codon pairs could possibly have evolutionary or survival advantages: upon a frameshift mutation, the peptide sequence of the frameshift would be more similar to the wild-type, and therefore, have better chances to remain/recover the wildtype function.

Here, we analyzed the shiftability of protein-coding genes in some model organisms. It will be interesting to investigate the shiftability of protein-coding genes in other species. Conceivably, the shiftability of protein-coding genes could possibly play an important role in the functioning, repairing and evolving of proteins and their coding genes. Finally, the frameshifts are not assumed to be tolerated in the sense that they are not changed. If they are not removed by selecting against, they are assumed to be preserved or repaired in the evolutionary history [27].

5. Conclusion

Through the above analysis, we conclude that the natural genetic code is nearly optimal in terms of shiftability (frameshift tolerability). The shiftability of coding genes guarantees a half-conservation of frameshifts, endows 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 for all species to tolerate frameshift mutations in both forward and backward directions, and thus, has a better fitness in the early evolution. Thanks to this ingenious property of the genetic code, the shiftability serves as an innate mechanism for cells to tolerate frameshift mutations occurred in protein-coding genes, by which the disasters of frameshift mutations were sometimes utilized as a driving force for molecular evolution.

Author Contributions

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

Acknowledgments

This study was funded by *National Natural Science Foundation of China* through Grant 31571369. We acknowledge Dr. Quanjiang Dong, Dr. Hervé Seligmanna, Prof. Michael Lynch, and other anonymous reviewers for their comments and proofreading of this article [16, 28].

Figure Legends

Fig 1. The alignment of the coding and the protein sequences of HIV/SIV GP120. (A) The alignment of GP120 coding sequences. Highlights: showing the frameshifting events. (B) The alignment of GP120 sequences, showing that the different GP120 sequences encoded in different reading frames of *gp120* are highly similar. The alignment was aligned by 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 compatible 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 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 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*	
1	*H. sapiens*	1573	2523	50	-3.06
2	*M. musculus*	1505	2591	190	-3.81
3	*X. tropicalis*	1660	2436	148	-3.80
4	*D. rerio*	1493	2603	148	-5.18
5	*D. melanogaster*	1418	2678	140	-5.02
6	*C. elegans*	1469	2627	164	-6.11
7	*A. thaliana*	1566	2530	15	-6.37
8	*S. cerevisiae*	1493	2603	159	-4.27
9	*E. coli*	1389	2707	197	-6.82
10	Equal Usage	0	0	0	-5.67

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

15. 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.

16. 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.

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

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

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

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

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

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

23. 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.

24. 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.

25. 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.

26. 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.

27. 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.

28. 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

```
                    *         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

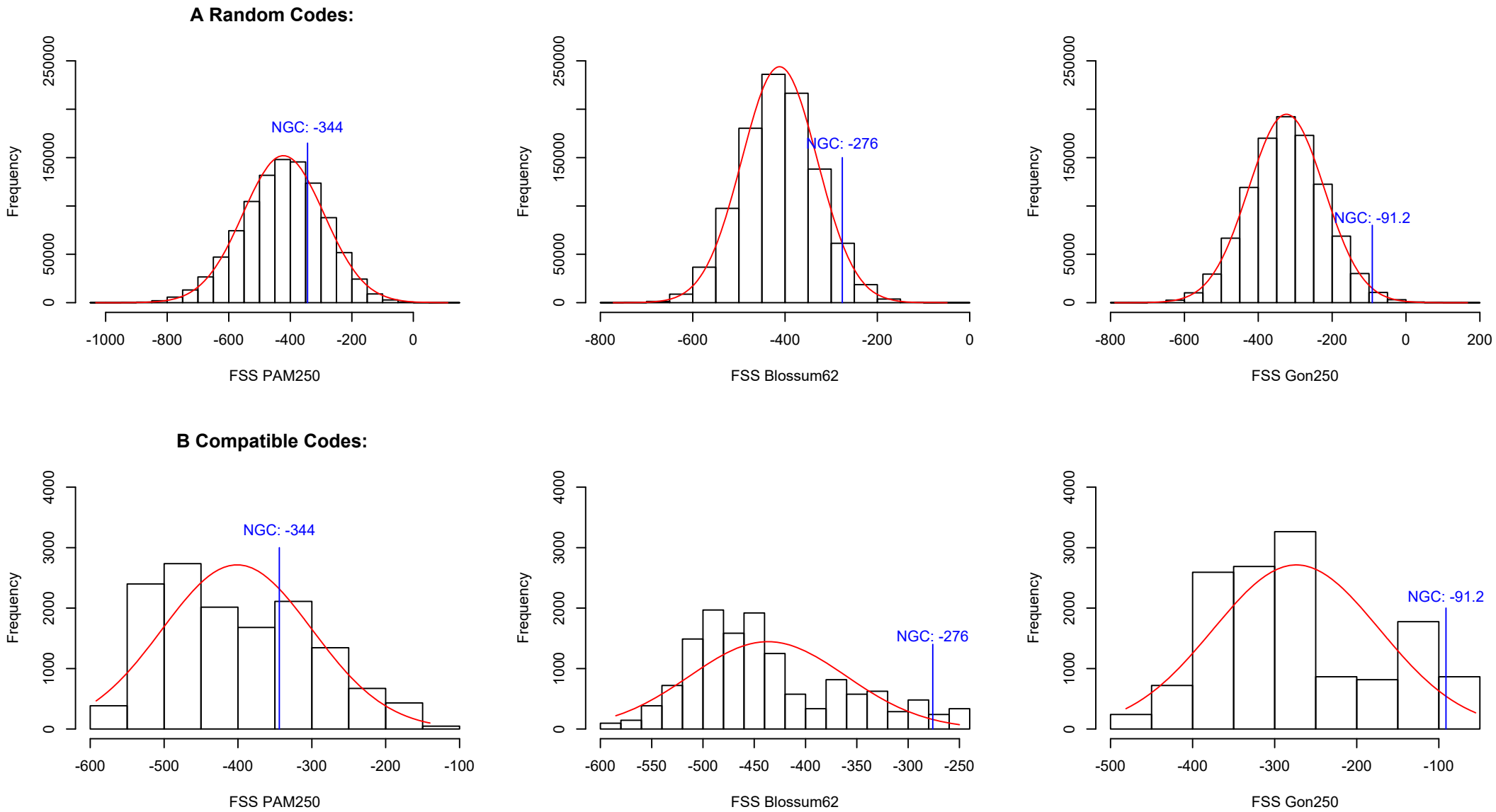
```
                        *        20         *
vegfaa     : MNLVVYLIQLFLAALLHLSAVKAAHIPKEGGKSKNDVI :  38
vegfaa-1   : MTWLFI*YSYFSRLSSICLL*RLPTYPKKGERAKMM*F :  35
vegfaa-1-r : MTWLFIWYSYFSRLSSICLLKRLPTYPKKGERAKMMWF :  38
vegfaa-2   : -MLGCLFDTVISRGSPPSVCCKGCPHTQRRGKEQK*CD :  36
vegfaa-2-r : -MLGCLFDTVISRGSPPSVCCKGCPHTQRRGKEQKWCD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