The Shiftability of Protein Coding Genes: The Genetic Code Was Optimized for Frameshift Tolerating

Xiaolong Wang[*1], *Xuxiang Wang, Gang Chen, Jianye Zhang, Yongqiang Liu, Chao Yang*

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

Abstract

The genetic code defines the relationship between a protein and its coding DNA sequence. It was presumed that most frameshifts would yield non-functional, truncated or cytotoxic products. In this study, we report that in *E. coli* a frameshift β-lactamase (bla) gene is still functional if all of the inner stop codons were readthrough or replaced by a sense codon. By analyzing a large dataset including all available protein coding genes in major model organisms, it is demonstrated that in any species, and in any protein-coding genes, the three translational products from the three different reading frames, are always similar to each other and with constant ~50% similarities and ~100% coverages, and the similarities is predefined by the genetic code rather than the sequences themselves, suggesting that the genetic code was optimized for frameshift tolerating in the early evolution, which endows every protein coding gene a *shiftability*, an inherent and everlasting ability to tolerate frameshift mutations, and serves as an innate mechanism for cells to deal with the frameshift problem. In addition, it is likely that every protein-coding gene can be translated into three isoforms from the three different reading frames, we proposed a new gene expression paradigm, "*one gene, three translations*", which is an amendment to the "*one gene, one/multiple peptides*" hypotheses.

[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.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

Introduction

The genetic code defines the relationship between the amino acid sequence of a protein and the DNA/mRNA sequence of the corresponding coding gene. The natural genetic code consist 64 triplet codons: 61 sense codons for specifying the 20 amino acids and the remaining three nonsense codons for stop signals.

Since the discovery of the genetic code [1], it has been revealed that the triplet codons have a number of interesting properties: (1) The genetic code are universal for all organisms, with a few small modifications in some organelle or organisms, such as mitochondrion and archaea; (2) The genetic code are redundant, degenerative and wobble (the third base tends to be interchangeable); (3) In an open reading frame, there is no punctuation exist between each pair of codons, so that frameshift mutations can be caused by an insertion or deletion (indel), while the reading frame is retained if the size of the indel is a multiple of three.

Although it has been reported that sometimes a partial frameshift is functional [2], a whole-frame shifting has been considered to be a completely loss of function (LOF), because not only every codon read and amino acid translated is changed, but often many nonsense codons are produced downstream the frameshift-causing indel. The "*Ambush Hypothesis*" [3] presumed that most frameshifts would yield non-functional proteins, lead to waste of energy, resources and activity of the biosynthetic machinery, and some peptides synthesized after frameshifts were thought to be cytotoxic [3-11]. Therefore, although it was observed that sometimes frame shifted or overlapped genes is functional [12-17], a frame-shifted translational product is generally considered to be non-functional, because it is a common sense that it is often possible to inactivate the function of a peptide by changing only one single residue.

On the other hand, it has been proved that the natural genetic code is optimized for translational error minimization [18], and thus is extremely efficient at minimizing the effects of point mutation or mistranslation errors [19]. In addition, because the frame-shifted codons for abundant amino acids overlap with the stop codons, hence

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

the robustness of the genetic code to frameshift errors is achieved by increasing the probability that a stop signal is encountered upon frameshift [20].

In this report, we demonstrated that the genetic code was optimized for frameshift tolerating, which endows every protein coding gene a characteristics of *shiftability*, an inherent ability to tolerate frameshift mutations. If all of the stop codons generated in a frameshift were readthrough or replaced by a sense codon, the translated frameshift isoform is highly similar to the original peptide, and might often still be functional.

Materials and Methods

1. Frameshift mutagenesis and back mutation

Using overlapping-extension polymerase chain reaction (OE-PCR), a technique for site-directed mutagenesis, a frameshift mutation of the *bla* gene is constructed by deleting one single nucleotide (G) in the upstream. A pair of mutagenesis primers was chemically synthesized by a commercial service provided by Shanghai Sangong Co. Ltd. The wild-type (*bla*+) and mutated (*bla*-) were cloned in the plasmid *pBR322* and transformed into *E. coli JM109*, grown on a tetracycline-containing plate (TCP) , the transformant colonies were picked up, propagated, and then plated on an ampicillin-containing plate (ACP). Fifty revertants were propagated in an ampicillin-containing broth (ACB), their plasmids were extracted, and their *bla* genes were sequenced by Sanger sequencing.

2. Alignment analysis of the frameshift isoforms

All available protein coding sequences in representative organisms, including *Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Xenopus tropicalis*, *Mus musculus* and *Homo sapiens*, were downloaded from the Ensemble Genome Database (Gene 78) using the BioMart data-mining tool. Ten thousand simulated protein coding sequences each containing 500 sense codons were generated by *Recodon* 1.6.0 [21] using default parameters. Frameshift mutations were constructed by deleting one or two bases in

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

the start codon, so that in the frameshift genes every codon is changed. All original and frameshift sequences were translated into protein sequences using the standard natural genetic code, but every stop codon that was generated in the frameshifting was readthrough by translating it into an amino acid according to Table 1. Multiple sequence alignment of the protein sequences and their frameshift isoforms were performed by ClustalW2. The pairwise similarity of the original peptide and a frameshift isoform is given by the percent of matched amino acid pairs that are similar (having a positive or zero amino acid substitution score).

Table 1. The natural correction tRNA for nonsense mutations in *E coli*.

Site	tRNA (AA)	Wild type		Correction	
		Code	Anti-code	Code	Anti-code
supD [22]	Ser (S)	→ UCG	CGA←	→ UAG	CUA←
supE [23]	Gln (Q)	→ CAG	CUG←	→ UAG	CUA←
supC	Tyr (Y)	→ UAC	GUA←	→ UAG	CUA←
supG [24]	Lys (K)	→ AAA	UUU←	→ UAA	UUA←
supU	Trp (W)	→ UGG	CCA←	→ UGA	UCA←

3. *Computational analysis of the codon substitutions*

A protein sequence consisting 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 (CDS) consists 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:* $B_{-1} B_0 \boldsymbol{B_1} \mid \boldsymbol{B_2 B_3} B_4 \mid B_5 B_6 \boldsymbol{B_7} \mid \ldots \mid \boldsymbol{B_{3i+2} B_{3i+3}} B_{3i+4} \mid B_{3i+5} B_{3i+6} \boldsymbol{B_{3i+7}} \mid \ldots$;

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

We can see that no matter how a frameshift is caused, 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 shifting* (*FF*): $\boldsymbol{B_3}\, B_4 B_5$ (A_{21});

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

And so does each of the downstream codons, produce two frameshift isoforms referred to as *FF* and *BF*. In either case, in every codon only one base is new, but in fact all three bases are changed when compared base by base with the original codon, so a frameshift substitution is actually different from a wobble or degenerative codon substitution. Traditionally, codon substitutions are classified into two types according to whether the encoded amino acid is changed or not: (1) *synonymous* (SS); (2) *nonsynonymous* (NSS). Based on above analysis, we classified codon substitutions into three subtypes: (1) *Random*; (2) *Wobble*; (3) *Frameshift* (FSS).

We wrote a java program, referred to as *Frameshift-CODON,* to compute the sum and average amino acid substitution scores for different kind of codon substitutions according to the standard genetic code and the substitution scoring matrices, including BLOSSUM62, PAM250 and GON250.

4. *Computational analysis of the codon pairs for frameshift substitutions*

For a given pair of amino acids, written as, $A_1 A_2$, where $A_i = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$, $i = 1, 2$; its encoding codon pair is written as, $\boldsymbol{B_1 B_2 B_3} \mid B_4 B_5 B_6$, where $B_k = \{A, G, U, C\}$, $k = 1 \ldots 6$. There are 400 different amino acid pairs and 4096 different codon pairs.

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

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

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

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

We wrote a java program referred to as *Frameshift-CODONPAIR* to compute the sum and average amino acid substitution scores for each kind of AA and codon pairs. The result of these calculations is a list of 400 AA pairs and their 4096 codon pairs, each with a frameshift substitution score (FSS).

5. *Computational analysis of the usage of codon and codon pairs*

The biased usage of codons and codon pairs was analyzed using the same method used in reference [25] on the same protein-coding sequences data used above. The program *CODPAIR* was rewritten in Java. For each sequence, it enumerates the total number of codons, the number of occurrences of each codon and each codon pair. The observed and expected frequency of each codon and dicodon is then calculated. The result of these calculations is a list of 64 codons and 3721 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 with the highest χ_1^2 value were identified as the most over- or under-represented dicodons, their frameshift substitution scores were computed, and compared with each other.

Results and Analysis

1. *Growth of E. coli with wild-type bla and the frameshift mutant*

As shown in Fig 1 and Fig 2A, when a plasmid *pBR322* containing a *bla*+ gene was transformed into *E. coli JM109,* the *bla*+ bacteria grow well on ACPs. When a frameshift mutation was introduced in the upstream of the *bla* gene, it was expected that there was no growth of the *bla*- bacteria on ACPs. However, repeatedly it was observed that there were always a few (about one out of 10^6~10^8) colonies that can grow on ACPs (Fig 2B). At first, we thought that these ampicillin-resistant colonies might be derived from a contamination of the wild-type bacteria. But no growth of the blank control (Fig 2C) suggested that the ampicillin-resistant colonies are not contamination of the wild-type bacteria.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

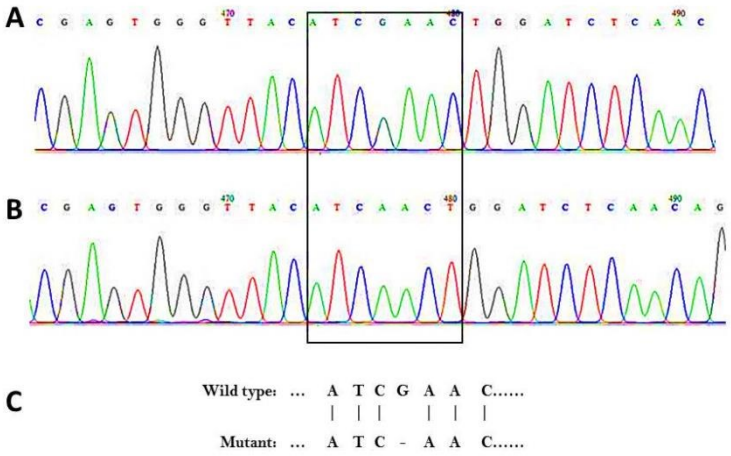

Fig 1. Introducing a frameshift mutation in the upstream of the *bla* gene in plasmid pBR322. Sanger sequencing result of (A) the wild type; (B) the frameshift mutant. (C) Alignment of the nucleotide sequence of the wild type and that of the frameshift mutant;

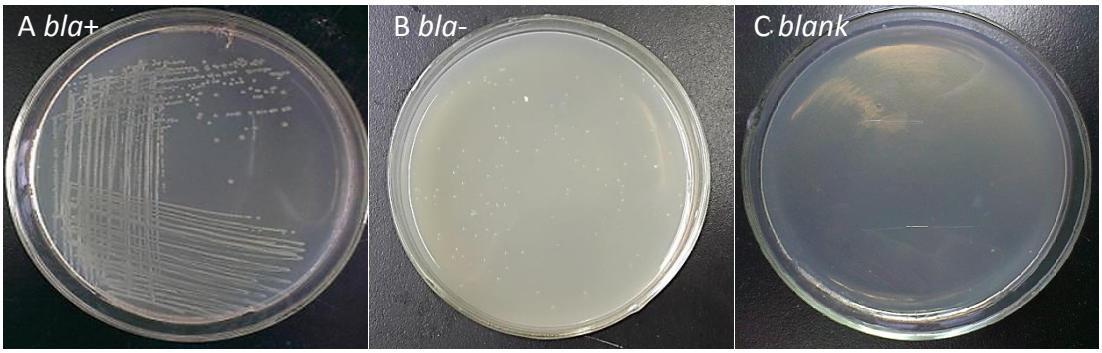

Fig 2. Growth of *E.coli JM109/pBR322* on ampicillin-containing LB plate: (A) Wild-type (*bla*+); (B) Frameshift (*bla*-); (C) Blank control.

Wild type	T	A	C	A	T	C	G	A	A	C	-	T	G	G	-	A	T	C	T	C	A	A	C	A	G	C	G	G	G
Mutant	T	A	C	A	T	C	~~G~~	A	A	C	-	T	G	G	-	A	T	C	T	C	A	A	C	A	G	C	G	G	G
Revertant 1	T	A	C	A	T	C	~~G~~	A	A	C	T	T	G	G	-	A	T	C	T	C	A	A	C	G	G	C	G	G	G
Revertant 2	T	A	C	A	T	C	~~G~~	A	A	C	C	T	G	G	-	A	T	C	T	C	A	A	C	G	G	C	G	G	G
Revertant 3	T	A	C	A	T	C	~~G~~	A	A	C	-	T	G	G	G	A	T	C	T	C	A	A	C	G	G	C	G	G	G

Fig 3. Sanger sequencing results, showing that the *bla* genes of the revertants are not the wild type, but different kind of back mutated frameshifts. Grey: shows the reading frame; Strikethrough: the base deleted in the mutagenesis; Red: the bases inserted in the revertants;

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

Sanger sequencing of the *bla* gene confirmed that they are not the wild type, but revertants. As shown in Fig 3, in the revertants the *bla* genes were repaired through different backward mutations, one different nucleotide was inserted downstream the base deleted in the mutagenesis, so that the reading frame was recovered downstream the insertion, so that only a few codons and their encoded amino acids between the deletion and the insertion were changed.

Upon until this point, it seems that there is nothing unusual, as it is a well-known phenomenon, *i.e.*, frameshift gene repairing or reading frame recovering (FGR/RFR) by a backward mutation, which was investigated as early as in the 1960s [2]. However, we felt that there is a logic contradiction: since a FGR/RFR must happen in a live cell, if a frameshift mutant itself could not survive in ACPs, how did the FGR/RFR happen in a cell that was dead? FGR/RFR is explained by a "*mutation-or-death*" model (Fig 4A): a backward mutation occurred naturally in the DNA replication process before the bacterial were killed. However, it is hard to believe that in the whole history of evolution, life have been betting their fates on such a high risk, because the rate of a naturally occurred backward mutation that had happened repaired a damaged gene might be even lower than that of a mutation that had damaged the same gene.

Therefore, in Fig. 3, the various independent backward mutations observed in the *bla* gene in the revertants are not the results of random backward mutations, but must be a targeted and programmed FGR/RFR. Obviously, the conquering of the problem of frameshift tolerating, frameshift gene repairing and reading frame recovering is extremely important for the existence of the species, and the underlying mechanism must be sophisticate, robust, target-oriented and well-controlled, and designed not for one individual gene, but for all genes in the genome as a whole.

Therefore, here we proposed a new "*readthrough-and-recovery*" model for FGR/RFR (Fig 4B): *Firstly*, the frameshift mutant itself is able to survive in ACP, because the frame-shifted gene is translated into a functional isoform by reading through the stop codons; *Secondly*, a bacteria cell "*knows*" which gene is frame-shifted, and it recruits a repairing machine to repair the damaged gene; *Thirdly*, the stop codons emerged in a frame-shifted gene or mRNA not only trigger the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

translational readthrough, but also serve as a signals for the repairing machine for the localization of the damaged gene, and then the reading frame is recovered by inserting a base in the upstream of the stop codons; *Finally*, the rate of reading through and reading frame recovery is slow, or the activity of the frameshift isoform is low, most of the cells were killed before the translational readthrough had happened, so that the survival rate is still very low. Nevertheless, this self-initiative model for the FGR/RFR process better explains the independent backward mutations when compared with the passive model relying solely on randomly occurred backward mutations.

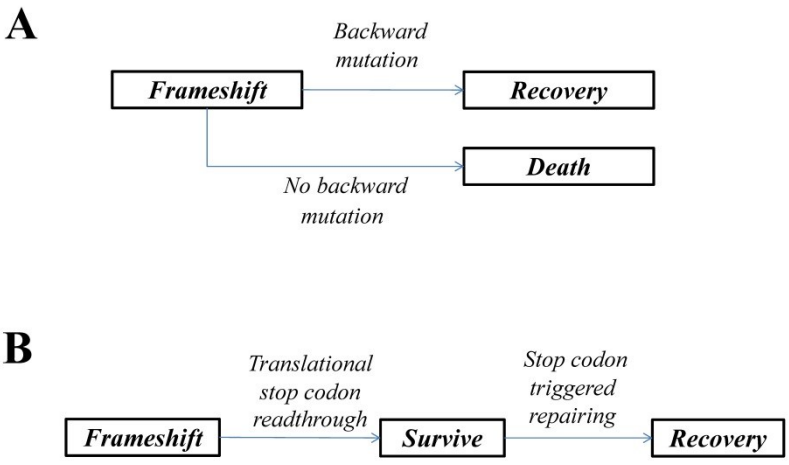

Fig 4. Two different models for frame-shifted gene repairing and reading frame recovery.

(A) The traditional "mutation-or-death" model; (B) This "readthrough-and-recovery" model.

2. *The frameshift isoforms are always highly similar to each other*

To find out the reason why the frame-shifted *bla-* gene is functional, the protein sequences of the wide type BLA and its two frameshift isoforms were aligned using ClustalW2. The alignment was displayed in GenDoc with the amino acids colored by their physiochemical properties (Fig 5). Surprisingly, both of the two frameshift isoforms are highly similar to the wild type BLA peptide, and most of the amino acids have similar physiochemical properties when compared with their aligned residues.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

As shown in Table 2, if the nonsense codons were ignored, on average 51% of the amino acids are conserved among the three isoforms, but there are 21 gaps in each sequence, caused mainly by the stop codons deleted in the frameshift CDSs. When every stop codon in the frame-shifted *bla-* was "*readthrough*" by translating it into an amino acid, an average of 45.8% of the sites remain conserved, and throughout the whole alignment in each sequence there are only 3 gaps, caused mainly by the bases deleted. Moreover, the similar amino acids distribute all over in the whole alignment, resulting in a near 100-percent coverage perfect alignment.

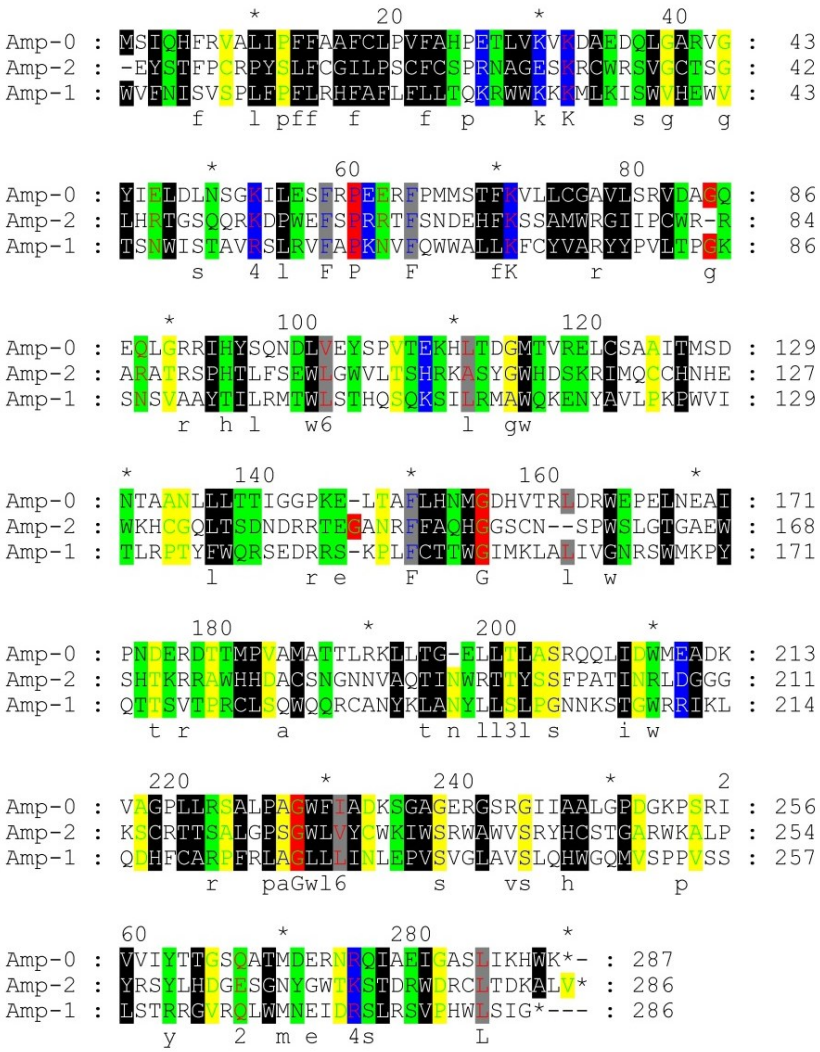

Fig 5. Alignment of the wide-type BLA and its frameshift isoforms. The alignment is aligned by ClustalW2, showed in GenDoc, and the amino acids were colored by their physiochemical properties. Every stop codon in the frameshift was translated according to Table 1.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

It has been observed that sometimes a frame-shifted protein coding gene is still functional [2, 14, 15, 26], but this phenomenon has been taken as special or individual cases, rather than a fundamentally important biological process sharing a common underlying mechanism. In fact, we found that this phenomenon is not rare, but can be observed quite often. For example, in different strains of HIV or SIV, such as HIV1J3, SIVCZ and SIVGB, quite a few forward and backward frameshifting events (Fig S1A, marked in yellow) occurred in both of the upstream and downstream of the envelop glycoproteins (GP120) genes, but their encoded protein sequences are highly similar (Fig S1B), and the frameshift isoforms are surely all functional. Since SIVGB is the ancestor of HIV1 and SIVCZ, it is imaginable that the origin of SIVCZ is caused by a number of frameshift events, and probably followed by a series of base substitutions which removed the stop codons generated.

Table 2. The alignment properties of BLA and its frameshift isoforms.

Readthrough	***Length***	***Number of Gaps***	***Similarity***			
			ORF-1-2	***ORF-1-3***	***ORF-2-3***	***Average***
Yes	291	21	0. 4914	0. 4639	0. 4192	0. 4582
No	287	3	0. 5052	0. 5366	0. 4983	0. 5134

In order to test whether or not this phenomenon is universal, we wrote a java program to align protein sequences with their frameshift isoforms on a large dataset, all available protein coding genes in the ensemble database in major model organisms, including *E. coli*, *S. cerevisiae*, *A. thaliana, C. elegans, D. melanogaster, D. rerio*, *X. tropicalis*, *M. musculus,* and *H. sapiens* and simulated. As shown in Table 3, in all of the natural and simulated sequences tested, the average pairwise similarity of the proteins and their frameshift isoforms, which was defined as the *shiftability* of the protein-coding genes, is centered approximately at 0.5. In other word, in any species, and for any protein coding genes, the amino acid sequences translated from the three different reading frames are always ca. 50% similar to each other. It is very likely that a coding gene/mRNA can be translated from each of the three different reading frames, the three translation products are three highly similar peptides, one main form and two

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

hidden frameshift isoforms. Therefore, as a supplement to the "*one gene, one/multiple peptides*" hypotheses, we proposed a new gene expression paradigm: "*one gene, three reading frames*", "*one peptide, three isoforms*", or "*one transcript, three translations*" (which one is the best? your choice!).

Table 3. The similarities of natural and simulated proteins and their frameshift isoforms.

NO	Species	Number of CDSs	Similarity			
			ORF-1-2	ORF-1-3	ORF-2-3	Average
1	*H. sapiens*	71857	0.5217±0.0114	0.5044±0.0122	0.4825±0.0147	0.5028±0.0128
2	*M. musculus*	4220	0.5180±0.0020	0.5011±0.0017	0.4801±0.0015	0.4997±0.0015
3	*X. tropicalis*	7706	0.5190±0.0013	0.4987±0.0013	0.4855±0.0008	0.5010±0.0008
4	*D. rerio*	14152	0.5162±0.0015	0.4921±0.0010	0.4901±0.0013	0.4995±0.0008
5	*D. melanogaster*	23936	0.5306±0.0007	0.5035±0.0008	0.5002±0.0010	0.5115±0.0006
6	*C. elegans*	29227	0.5210±0.1379	0.4813±0.0015	0.5073±0.0010	0.5032±0.0461
7	*A. thaliana*	35377	0.5389±0.0508	0.5078±0.0481	0.5062±0.0480	0.5176±0.0388
8	*S. cerevisiae*	5889	0.5234±0.0007	0.5022±0.0008	0.4921±0.0005	0.5059±0.0004
9	*E.coli*	4140	0.5138±0.0019	0.4871±0.0046	0.4810±0.0015	0.4940±0.0012
10	*Simulated*	10000	0.5165±0.0282	0.4745±0.0272	0.4773±0.0263	0.4894±0.0013

3. *The genetic code was optimized for frameshift tolerating*

As shown in Table 3, the similarities among a protein and its frameshift isoforms are similar in all species, and the standard deviation is very small, suggesting that the shiftability is largely independent on the species and the DNA or protein sequences, implying that the shiftability is defined by the genetic code rather than the sequence of the proteins or their coding sequences.

As described above in the method section, we computed the average amino acid substitution scores for different kind of codon substitutions, including random, wobble, forward and backward frameshift substitutions. As shown in Table 4, in all 4096 possible codon substitutions, except for the 64 unchanged codons, only a small proportion (4.1%) of the 4032 changed codons are synonymous and the other 95.9% are nonsynonymous; in addition, 80% (128/166) of the synonymous substitutions are wobble, and 66.7% (128/192) of the wobble substitutions are synonymous and the other 33.3% are nonsynonymous, and therefore the average substitution score for the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

wobble substitutions is the highest. In contrast, 95% of the frameshift substitutions are nonsynonymous and only 5% of them are synonymous (Table 5). In addition, 21% of the random substitutions are positive nonsynonymous, and only 15.6% of the wobble substitutions are positive nonsynonymous, while as many as (72+76)/512=28.9% of the frameshift substitutions are positive nonsynonymous, which is much higher than the proportion of positive NSSs in the other groups. Obviously, in the natural genetic code, wobble substitutions are designed mainly for synonymous substitutions, while frameshift substitutions are assigned mainly to conserved amino acid substitutions.

Table 4. The amino acid substitution scores for different kind of codon substitutions.

Codon Substitution		*ALL*	*Frameshift*		*Wobble*	*Others*
		(Random)	***FF***	***BF***		
Number of Substitutions	*All*	4096	256	256	256	3328
	Unchanged (%)	64 (1.6%)	4 (1.6%)	4 (1.6%)	64 (25%)	0
	Changed (%)	4032 (98.4%)	252 (98.4%)	252 (98.4%)	192 (75%)	3328 (100%)
	SS (%)	166 (4.1%)	14 (5.5%)	14 (5.5%)	128 (50%)	2 (0.1%)
	NSS-Positive (%)	859 (21%)	76 (29.7%)	72 (28.1%)	40 (15.6%)	671 (20.2%)
	NSS-Negative (%)	3007 (73.4%)	162 (63.3%)	166 (64.8%)	24 (9.4%)	2655 (79.8%)
Average Substitution Score	*BLOSSUM62*	-1.29	-0.61	-0.65	3.77	-1.34
	PAM250	-1.75	-0.84	-0.84	3.68	-1.81
	GON250	-10.81	-2.84	-2.84	36.13	-11.34

Table 5. The synonymous frameshift substitutions and their amino acid substitution scores.

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

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

In addition, no matter which amino acid substitution scoring matrix was used for computation, the average substitution score of the frameshift substitutions are always significantly higher than those of the random substitutions ($P < 0.01$), suggesting that more similar amino acid pairs are selected for frameshift substitutions when compared with random substitutions. Therefore, the similarities among protein sequences and their frameshift isoforms are predefined by the genetic code and independent on the proteins or their coding sequences themselves. Because of the frameshift substitutions, in addition to the degenerative codons, it is guaranteed that in any protein nearly half of the amino acids in a frameshift isoform are changed into similar residues, which explained the observed 50% similarities and 100% coverage among the three isoforms among all species.

4. The shiftability at sequence level

Although the code-level shiftability guaranteed a 50% similarity among a protein and its frameshift isoforms, it does not necessarily imply that all frameshift isoforms of a protein have a function. However, it does form the basis for the toleration of whole-frame or partial frameshifts. In addition, the other 50% of sites are changed into less similar amino acids, also provides a basis for molecular evolution, such as structural and functional improvements of the protein, overlapping genes, and so on.

Although the shiftability of a coding sequence is predefined mainly by the genetic code, an additional shiftability might be able to be maintained at a sequence level. We thought that a functionally important coding gene which is more conserved, such as a housekeeping gene, might has higher shiftability when compared with a variable non-housekeeping gene. At first, we thought that a biased usage of codons may contribute to the sequence-level shiftability. However, it is somewhat surprising that the average FSSs of a biased usage of the codons are even worse when compared with that of an equal usage of them (Table 6), but the difference is not significant, suggesting that the biases usage of codons does not contribute directly to the shiftability of the genes, but they may have an indirect impact on the shiftability, for example, by shaping the pattern of codon pairs.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

Given a pair of amino acids, $A_1 A_2$, if A_1 and A_2 have, respectively, m_1 and m_2 degenerative codons, their encoding codon pair, $B_1 B_2 B_3 \mid \boldsymbol{B_4 B_5 B_6}$, has $m_1 \times m_2$ possible combinations, called *degenerative codon pairs* (DCPs). It has been widely reported that the usage of codon pairs are highly biased in various species, including virus, bacteria, human and animals [25, 27-32]. However, as shown in Table 7, in human, the average FSSs of the most over- and under-represented dicodons are not significantly different from each other, *i.e.*, DCPs that are more frameshift tolerating are not used more frequently, suggesting that the shiftability is independent on the usage of codons and codon pairs. Therefore, the shiftability at the sequence level, if exist, is not achieved universally by biased usages of codons and dicodons, but in a more complicated or gene-specific model.

Table 6. The usage of codons in *E. coli* and their frameshift substitution scores (GON250)

1st	Codons				3rd	Usage (%)				FSS (FF+BF)				Biased Usage (FSS*Usage)				Equal Usage FSS*100/64			
U	UUU	UCU	UAU	UGU	U	1.90	1.10	1.60	0.40	281	-161	-41	-39	533.9	-177.1	-65.6	-15.6	439.1	-251.6	-64.1	-60.9
	UUC	UCC	UAC	UGC	C	1.80	1.00	1.40	0.60	-11	-61	-94	-19	-19.8	-61.0	-131.6	-11.4	-17.2	-95.3	-146.9	-29.7
	UUA	UCA	UAA	UGA	A	1.00	0.70	0.20	0.10	106	-77	0	0	106.0	-53.9	0.0	0.0	165.6	-120.3	0.0	0.0
	UUG	UCG	UAG	UGG	G	1.10	0.80	0.03	1.40	69	-85	0	-210	75.9	-68.0	0.0	-294.0	107.8	-132.8	0.0	-328.1
C	CUU	CCU	CAU	CGU	U	1.00	0.70	1.20	2.40	51	-8	-103	-99	51.0	-5.6	-123.6	-237.6	79.7	-12.5	-160.9	-154.7
	CUC	CCC	CAC	CGC	C	0.90	0.40	1.10	2.20	-153	388	-36	-43	-137.7	155.2	-39.6	-94.6	-239.1	606.3	-56.3	-67.2
	CUA	CCA	CAA	CGA	A	0.30	0.80	1.30	0.30	-69	58	42	-17	-20.7	46.4	54.6	-5.1	-107.8	90.6	65.6	-26.6
	CUG	CCG	CAG	CGG	G	5.20	2.40	2.90	0.50	-106	48	32	-59	-551.2	115.2	92.8	-29.5	-165.6	75.0	50.0	-92.2
A	AUU	ACU	AAU	AGU	U	2.70	1.20	1.60	0.70	-19	-69	-82	-35	-51.3	-82.8	-131.2	-24.5	-29.7	-107.8	-128.1	-54.7
	AUC	ACC	AAC	AGC	C	2.70	2.40	2.60	1.50	-167	-13	44	49	-450.9	-31.2	114.4	73.5	-260.9	-20.3	68.8	76.6
	AUA	ACA	AAA	AGA	A	0.40	0.10	3.80	0.20	-109	-23	139	48	-43.6	-2.3	528.2	9.6	-170.3	-35.9	217.2	75.0
	AUG	ACG	AAG	AGG	G	2.60	1.30	1.20	0.20	-95	-25	115	6	-247.0	-32.5	138.0	1.2	-148.4	-39.1	179.7	9.4
G	GUU	GCU	GAU	GGU	U	2.00	1.80	3.30	2.80	-25	-33	-149	-126	-50.0	-59.4	-491.7	-352.8	-39.1	-51.6	-232.8	-196.9
	GUC	GCC	GAC	GGC	C	1.40	2.30	2.30	3.00	-103	27	-5	26	-144.2	62.1	-11.5	78.0	-160.9	42.2	-7.8	40.6
	GUA	GCA	GAA	GGA	A	1.20	2.10	4.40	0.70	-85	-5	42	-8	-102.0	-10.5	184.8	-5.6	-132.8	-7.8	65.6	-12.5
	GUG	GCG	GAG	GGG	G	2.40	3.20	1.90	0.90	-89	-9	12	270	-213.6	-28.8	22.8	243.0	-139.1	-14.1	18.8	421.9
2nd	U	C	A	G		Average: 1.56				Average: -14.25				Average: -31.63				Average: -22.27			

Table 7. FSSs for the most over- and under-represented dicodons in human.

A. Most over-represented dicodons												B. Most under-represented dicodons											
1	2	AA1	AA2	FSS1	FSS2	FF	BF	AA-FF	AA-BF	FSS-FF	FSS-BF	1	2	AA1	AA2	FSS1	FSS2	FF	BF	AA-FF	AA-BF	FSS-FF	FSS-BF
GCG	GCG	A	A	-9	-9	CGG	GGC	R	G	-6	5	GUC	GAA	V	E	-103	42	UCG	CGA	S	R	-10	4
CCG	CCG	P	P	48	48	CGC	GCC	R	A	-9	3	CUC	GAA	L	E	-153	42	UCG	CGA	S	R	-21	4
CGC	UGU	R	C	-43	-39	GCU	CUG	A	L	-6	-15	CGC	GAA	R	E	-43	42	GCG	CGA	A	R	-6	4
CGC	UGC	R	C	-43	-19	GCU	CUG	A	L	-6	-15	GUC	GAG	V	E	-103	12	UCG	CGA	S	R	-10	4
UGU	GGG	C	G	-39	270	GUG	UGG	V	W	0	-40	UUC	GAA	F	E	-11	42	UCG	CGA	S	R	-28	4
CUU	CGA	L	R	51	-17	UUC	UCG	F	S	20	-2	UUC	GUA	F	V	-11	-85	UCG	CGU	S	R	-28	-20
CGC	UGG	R	W	-43	-210	GCU	CUG	A	L	-6	-7	CUC	GCA	L	A	-153	-5	UCG	CGC	S	R	-21	-6
AGU	GGG	S	G	-35	270	GUG	UGG	V	W	-10	-40	CGC	GAU	R	D	-43	-149	GCG	CGA	A	R	-6	-3
GUC	ACC	V	T	-103	-13	UCA	CAC	S	H	-10	-3	GGC	GAA	G	E	26	42	GCG	CGA	A	R	5	4
UCC	UCG	S	S	-61	-85	CCU	CUC	P	L	4	-21	GCC	GAU	A	D	27	-149	CCG	CGA	P	R	3	-3
UGU	GAC	C	D	-39	-5	GUG	UGA	V	*	0	-3	GGU	AAG	G	K	-126	115	GUA	UAA	V	*	-33	-4
AAU	GGG	N	G	-82	270	AUG	UGG	M	W	-22	-40	UUC	GCA	F	A	-11	-5	UCG	CGC	S	R	-28	-6
AGC	AGC	S	S	49	49	GCA	CAG	A	Q	11	2	GUC	GCA	V	A	-103	-5	UCG	CGC	S	R	-10	-6
GUC	AUC	V	I	-103	-167	UCA	CAU	S	H	-10	-22	CUC	GGA	L	G	-153	-8	UCG	CGG	S	R	-21	-10
ACC	AUC	T	I	-13	-167	CCA	CAU	P	H	1	-22	UUC	GCU	F	A	-11	-33	UCG	CGC	S	R	-28	-6
GUC	ACU	V	T	-103	-69	UCA	CAC	S	H	-10	-3	UGC	GCA	C	A	-19	-5	GCG	CGC	A	R	5	-6
CUC	UUC	L	F	-153	-11	UCU	CUU	S	L	-21	20	GCC	GAA	A	E	27	42	CCG	CGA	P	R	3	4
UGU	GGC	C	G	-39	26	GUG	UGG	V	W	0	-40	UCC	GUA	S	V	-61	-85	CCG	CGU	P	R	4	-20
UCG	GCG	S	A	-85	-9	CGG	GGC	R	G	-2	5	UCC	GCU	S	A	-61	-33	CCG	CGC	P	R	4	-6
GGU	GUC	G	V	-126	-103	GUG	UGU	V	C	-33	0	AGC	GAA	S	E	49	42	GCG	CGA	A	R	11	4
Average				-24.025		Average				-8.825						-29.425						-6.975	
																T-test: p= 0.79						T-test: p= 0.54	

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

Discussion

1. Frameshift tolerating and the ambush hypothesis

Frameshift events have been thought to be a waste of energy and resources, and frameshift peptide products were thought to have unpredictable cytotoxic effects [8]. The *ambush hypothesis* suggested there is a selective pressure favoring the evolution of hidden (out-of-frame) stop codons [5-7]. It was showed that hidden stops have been evolved under positive selection for the minimization of frame-shifted errors [6]. However, although this hypothesis gained some support in whole-genome studies, it is limited at a single-gene scale. For example, the polyketide synthase (PKS) genes presented with significantly lower level of hidden stop codons than expected, suggesting both non-adherence to the ambush hypothesis and a suppression of hidden stop codon evolution [8]. In addition, it was reported that some sense codons have a more significant excess than stop codons [4]. These controversial results can be well explained if the emerging of the hidden stops after the occurrence of a frameshift is considered to be a signal to trigger the cell machine for readthrough and then for reading frame recovery, rather than a signal for translational termination. Because the frame-shifted translation products are not wastes but useful, thus a moderate or low level of stop signal is enough for triggering readthrough and reading frame recovery, therefore the hidden stop codons are not necessarily to have an excess in every gene.

2. Shiftability and the pseudogenes

In addition, a large number of "*pseudogenes*", containing usually a frameshift or nonsense mutations, exist widely in the genome of various species, including bacteria, yeast [33], human and animals [14, 15]. Although they are considered non-functional and most of them would be removed finally in the evolution process, some of them may be functional, or serve as a backup to the main functional gene, and play a role in the functioning and evolution of proteins and their coding genes. It has been reported that in *E. coli* the levels of stop codon readthrough and frameshifting are both high and growth phase dependent [34], so it is possible that a pseudogene may sometimes

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

be functional. If so, pseudogenes can be seen as "reading frame transitional" genes, which is mutating under a selection pressure, which will finally be removed or turned into a functional form when the inner stop codons were replaced by sense codons.

3. The universality of the shiftability

Here we experimentally validated the shiftability of a protein-coding gene only in *E. coli*, thus it is interesting to ask whether or not the mechanism is preserved in other species. It has been reported that in some animal species frameshift mutations in mitochondrial genes are tolerated by the translation systems [13, 16, 17]. For example, a +1 frameshift insertion has been tolerated in the *nad3* in some birds and reptiles [13]. Moreover, frame-shifted overlapping genes have been found in mitochondria genes in fruit fly and turtles [35, 36]. Meanwhile, translational stop codon readthrough has been widely observed in virus, bacteria and many other species [33, 37-43]. Although frameshift tolerating observed in these species has been explained by the *programed translational frameshifting* mechanism [12, 44-46], with the satisfaction of several prerequisite conditions, such as the constant shiftability, the frameshift tolerating and the translational stop codon readthrough, has been widely observed in many species, it is very likely that the shiftability works and contributes, at least partially, to the expression, functioning, repairing and evolution of protein coding genes in all species.

Conclusion

The natural genetic code is a result of selection in the early evolution, and it has a number of superiorities when compared with the other possible genetic codes [47-58]. It was pointed out that the natural genetic code is optimized for translational error minimization, because amino acids whose codons differed by a single base in the first and third codon positions were similar with respect to polarity and hydropathy, and the differences between amino acids were specified by the second codon position is most easily explained by selection to minimize deleterious effects of translation errors during the early evolution of the genetic code [18]. In addition, it was proved that

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

only one in every million alternatives is more efficient than the natural genetic code, which is extremely efficient at minimizing the effects of point mutation or translation errors [19]. Recently, it was reported that the natural genetic code is nearly optimal for allowing additional information within coding sequences [20].

We here showed that the standard genetic code keep an amino acid unchanged for degenerative substitutions, similar for frameshift substitutions, and different for other non-degenerative and non-frameshift substitutions. Based on the above experimental, theoretic and data analysis, we concluded that the natural genetic code was optimized for frameshift tolerating. The ingenious "*underlying design*" of the natural codon table endows every protein-coding gene a constant shiftability, an inherent and everlasting ability to tolerate frameshift mutations, and endows the owner creatures a powerful ability that can survive on frameshift mutations in any coding gene, and thus be highly superior and win the competence of survival in the early evolution, and finally became the universal genetic code of choice by all creatures. Conceivably, the shiftability of the protein-coding genes is fundamentally important for the survival, competence, adaption and evolution of species. It serves as an innate mechanism for cell to deal with the frameshift problem, which might exist from the beginning of, or even before, the origin of life.

Author Contributions

XLW conceived the main idea, designed the experiments, coded the programs, analyzed the data and wrote the paper; the other authors performed the experiments, and discussion on the paper.

Acknowledgements:

This research was supported by National Natural Science Foundation of China (Grants 81072567) and Shandong Provincial Natural Science Foundation (Grant ZR2010HM056). We appreciate Dr. Xingguo Liang from Ocean University of China and Dr. Quanjiang Dong from Qingdao Municipal Hospital for helpful discussions.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

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. Streisinger, G., et al., *Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday.* Cold Spring Harb Symp Quant Biol, 1966. **31**: p. 77-84.
3. 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.
4. Morgens, D.W., C.H. Chang, and A.R.O. Cavalcanti, *Ambushing the ambush hypothesis: predicting and evaluating off-frame codon frequencies in Prokaryotic Genomes.* Bmc Genomics, 2013. **14**.
5. Seligmann, H., *The ambush hypothesis at the whole-organism level: Off frame, 'hidden' stops in vertebrate mitochondrial genes increase developmental stability.* Computational Biology and Chemistry, 2010. **34**(2): p. 80-85.
6. Singh, T.R. and K.R. Pardasani, *Ambush hypothesis revisited: Evidences for phylogenetic trends.* Computational Biology and Chemistry, 2009. **33**(3): p. 239-244.
7. Seligmann, H. and D.D. Pollock, *The ambush hypothesis: Hidden stop Ccodons prevent off-frame gene reading.* DNA and Cell Biology, 2004. **23**(10): p. 701-705.
8. Bertrand, R.L., M. Abdel-Hameed, and J.L. Sorensen, *Limitations of the 'ambush hypothesis' at the single-gene scale: what codon biases are to blame?* Mol Genet Genomics, 2014.
9. Morgens, D.W., C.H. Chang, and A.R. Cavalcanti, *Ambushing the Ambush Hypothesis: predicting and evaluating off-frame codon frequencies in prokaryotic genomes.* BMC Genomics, 2013. **14**: p. 418.
10. Seligmann, H., *The ambush hypothesis at the whole-organism level: Off frame, 'hidden' stops in vertebrate mitochondrial genes increase developmental stability.* Comput Biol Chem, 2010. **34**(2): p. 80-5.
11. Singh, T.R. and K.R. Pardasani, *Ambush hypothesis revisited: Evidences for phylogenetic trends.* Comput Biol Chem, 2009. **33**(3): p. 239-44.
12. Farabaugh, P.J., *Programmed translational frameshifting.* Microbiological Reviews, 1996. **60**(1): p. 103-&.
13. 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.
14. 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.* Journal of Biological Chemistry, 2004. **279**(26): p. 27383-27389.
15. Baykal, U., A.L. Moyne, and S. Tuzun, *A frameshift in the coding region of a novel tomato class I basic chitinase gene makes it a pseudogene with a functional wound-responsive promoter.* Gene, 2006. **376**(1): p. 37-46.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

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. 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.

19. 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.

20. 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.

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

22. 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.

23. 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.

24. 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.

25. 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.

26. Xu, J., R.W. Hendrix, and R.L. Duda, *Conserved translational frameshift in dsDNA bacteriophage tail assembly genes.* Molecular Cell, 2004. **16**(1): p. 11-21.

27. Das, G. and R.H.D. Lyngdoh, *Configuration of wobble base pairs having pyrimidines as anticodon wobble bases: significance for codon degeneracy.* Journal of Biomolecular Structure & Dynamics, 2014. **32**(9): p. 1500-1520.

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

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

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

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

32. 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.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

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

34. 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.

35. 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.

36. 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.

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

38. 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).

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

40. Howard, M.T., et al., *Readthrough of dystrophin stop codon mutations induced by aminoglycosides.* Annals of Neurology, 2004. **55**(3): p. 422-426.

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

42. 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.

43. 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.

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

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

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

47. 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.

48. 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.

49. 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.

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

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

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015

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

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

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

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

56. 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.

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

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

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.806v1 | CC-BY 4.0 Open Access | rec: 23 Jan 2015, publ: 23 Jan 2015