Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 337 (2005) 692–700
BBRC
www.elsevier.com/locate/ybbrc

Determination of mutation trend in proteins by means of translation probability between RNA codes and mutated amino acids

Guang Wu *, Shaomin Yan

Computational Mutation Project, DreamSciTech Consulting, 301, Building 12, Nanyou A-zone, Jiannan Road, Shenzhen, Guangdong Province CN-518054, China

Received 5 September 2005
Available online 26 September 2005

Abstract

In this study, we estimate the translation probability to amino acid from RNA codon. With the determined 183 translation probabilities and amino-acid composition of eight highly mutated proteins, we construct the theoretical distributions of mutated amino acids in these proteins and then compare them with their actual distributions affected by mutations. Thereafter we trace the pattern of translation probabilities from RNA codons to mutated amino acids of 1053 point missense mutations. Finally, we statistically conclude that the natural mutation trend goes along the theoretical translation probability.
© 2005 Elsevier Inc. All rights reserved.

Keywords: Amino acid; Mutation; RNA; Translation probability

Amino-acid coding obeys the standard genetic codes: four distinct symbols grouped in clusters of three elements, or triplets, code for the 20 different amino acids and the STOP signal which marks the end of protein transcription [1]. Translation of the genetic code requires attachment of tRNAs to their cognate amino acids, and the editing by phenylalanyl-tRNA synthetase is essential for faithful translation of the genetic code [2]. Translation could be enhanced by increasing the rate of elongation, reducing the cost of proofreading [3,4], increasing the accuracy of translation [5–7], or by any combination of those mechanisms [8]. Also protein charge heterogeneity [9] and protein release factors [10] can influence the translation of mRNAs.

Genetic mutation engineers the mutation at the protein level. Between RNA and protein, the RNA codes have the unambiguous relationship with their translated amino acids, for example, any of four RNA codons ACU, ACC, ACA, and ACG can be translated to the amino acid threonine. Between DNA and RNA, a single-base change in DNA nucleotide leads to the corresponding change in the RNA code. For instance, the RNA codon ACU can be changed to ACC, ACA, ACG, AUU, AAU, AGU, UCU, CCU, and GCU due to a single-base change in DNA, if we do not consider the possibility that U changes to U in the RNA codon ACU. As a result, each change in RNA code will be translated to different amino acids (Table 1). Table 1 shows that the translated amino acids induced by the changes in RNA code are not equally distributed (the last two rows).

This provides us a way to estimate the mutation trend from RNA to amino acid. For example, we have a RNA sequence and its protein sequence, we are interested in the RNA codon ACU and we would like to know which type of mutated amino acid is likely to be formed by an unspecified change in RNA codon ACU. From Table 1 we know that the mutated amino acid is highly likely to be threonine because it has the largest probability whereas other types of mutated amino acids have small probabilities. This is very suggestive because we can eventually define the mutated amino acids with the largest translation probability based on the change in RNA codes.

On the other hand, the explicit corresponding relationship between RNA codons and their translated amino

* Corresponding author. Fax: +86 755 2528 8156.
E-mail address: hongguanglishibahao@yahoo.com (G. Wu).

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.09.106

Table 1
Possible changes in the RNA codon ACU and the related changes in the translated amino acids

RNA codon				Translated amino acid
First position	Second position	Third position		
Change in the first position				
A	C	U	→	Threonine
U	C	U	→	Serine
C	C	U	→	Proline
G	C	U	→	Alanine
	Change in the second position			
A	C	U	→	Threonine
A	U	U	→	Isoleucine
A	A	U	→	Asparagine
A	G	U	→	Serine
		Change in the third position		
A	C	U	→	Threonine
A	C	C	→	Threonine
A	C	A	→	Threonine
A	C	G	→	Threonine
Translated amino acids	1 alanine + 1 asparagine + 1 isoleucine + 1 proline + 2 serines + 6 threonines			
Translation probability	1/12 + 1/12 + 1/12 + 1/12 + 2/12 + 6/12			

acids with different translation probabilities provides us another way to estimate the mutation trend from amino acid to RNA. For example, we still have a RNA sequence and its protein sequence with many documented mutations. We are interested in the amino acid threonine and its mutated amino acids, and we would like to know the distribution of translation probability from RNA codons to threonine and its mutated amino acids in order to analyze the mutation trend from RNA to amino acid.

The amino acid threonine can be translated from four RNA codons ACU, ACC, ACA, and ACG. Naturally, we can have the translation probabilities with respect to whether or not we include the cases of "U" changing to "U," "A" changing "A," "C" changing to "C", and "G" changing "G" (Table 2). As can be seen in Table 2, threonine can be mutated to eight different amino acids with different translation probabilities if we do not calculate the probabilities that "U" changes to "U", "A" changes to "A," "C" changes to "C", and "G" changes to "G." So far, we have yet to know the probabilistic pattern in the proteins with many mutations regarding the changes in the corresponding RNA codons, intuitively we might not expect the mutations to follow the maximal translation probability (the last two rows in Table 2).

Comparing these two approaches offered by the translation probability, the first one can be used to predict what will happen in future, i.e., which type of amino acid will be likely to be formed in future from a change in RNA codon such as those in Table 1. The second one can be used to analyze what happened in the past, that is, which RNA code was changed and led to the mutation at amino acid level such as those in Table 2.

Table 2
Mutated amino acids and their translation probabilities with regard to the changes in four RNA codons ACU, ACC, ACA, and ACG

RNA codon	ACU	ACC	ACA	ACG
Original amino acid	Threonine	Threonine	Threonine	Threonine
Mutated amino acids				
Changes in the first position in RNA codon	Serine	Serine	Serine	Serine
	Proline	Proline	Proline	Proline
	Alanine	Alanine	Alanine	Alanine
Changes in the second position in RNA codon	Isoleucine	Isoleucine	Isoleucine	Methionine
	Asparagine	Asparagine	Lysine	Lysine
	Serine	Serine	Arginine	Arginine
Changes in the third position in RNA codon	Threonine	Threonine	Threonine	Threonine
	Threonine	Threonine	Threonine	Threonine
	Threonine	Threonine	Threonine	Threonine
Total I	4 alanines + 2 arginines + 2 asparagines + 3 isoleucines + 2 lysines + methionine + 4 prolines + 6 serines + 12 threonines			
Translation probability I	4/36 + 2/36 + 2/36 + 3/36 + 2/36 + 1/36 + 4/36 + 6/36 + 12/36			
Total II	4 alanines + 2 arginines + 2 asparagines + 3 isoleucines + 2 lysines + methionine + 4 prolines + 6 serines			
Translation probability II	4/24 + 2/24 + 2/24 + 3/24 + 2/24 + 1/24 + 4/24 + 6/24			

I and II indicate the inclusion and exclusion, respectively, of the same type of amino acids before and after mutation, i.e., the mutated amino acid is threonine.

Table 3
Mutated amino acids and their translation probabilities

Original amino acid	RNA codon	Mutated amino acids translated from the changed RNA codon			Number of translated amino acids with their translation probability
		First position	Second position	Third position	
Phe	UUU	Leu, Ile, Val	Ser, Tyr, Cys	Phe, Leu, Leu	2Cys + 2Ile + 6Leu + 2Ser + 2Val + 2Tyr
	UUC	Leu, Ile, Val	Ser, Tyr, Cys	Phe, Leu, Leu	2/16 + 2/16 + 6/16 + 2/16 + 2/16 + 2/16
Leu	UUA	Leu, Ile, Val	Ser, STOP, STOP	Phe, Phe, Leu	4Phe + Ile + Met + 2Ser + 2Val + Trp + 3STOP
	UUG	Leu, Met, Val	Ser, Trp, STOP	Phe, Phe, Leu	4/14 + 1/14 + 1/14 + 2/14 + 2/14 + 1/14 + 3/14
Leu	CUU	Phe, Ile, Val	Pro, His, Arg	Leu, Leu, Leu	2Phe + 2His + 3Ile + Met + 4Pro + 2Gln + 4Arg + 4Val
	CUC	Phe, Ile, Val	Pro, His, Arg	Leu, Leu, Leu	2/22 + 2/22 + 3/22 + 1/22 + 4/22 + 2/22 + 4/22 + 4/22
	CUA	Leu, Ile, Val	Pro, Gln, Arg	Leu, Leu, Leu	
	CUG	Leu, Met, Val	Pro, Gln, Arg	Leu, Leu, Leu	
Ile	AUU	Phe, Leu, Val	Thr, Asn, Ser	Ile, Ile, Met	2Phe + Lys + 4Leu + 3Met + 2Asn + Arg + 2Ser + 3Thr + 3Val
	AUC	Phe, Leu, Val	Thr, Asn, Ser	Ile, Ile, Met	2/21 + 1/21 + 4/21 + 3/21 + 2/21 + 1/21 + 2/21 + 3/21 + 3/21
	AUA	Leu, Leu, Val	Thr, Lys, Arg	Ile, Ile, Met	
Met	AUG	Leu, Leu, Val	Thr, Lys, Arg	Ile, Ile, Ile	3Ile + Lys + 2Leu + Arg + Thr + Val 3/9 + 1/9 + 2/9 + 1/9 + 1/9 + 1/9
Val	GUU	Phe, Leu, Ile	Ala, Asp, Gly	Val, Val, Val	4Ala + 2Asp + 2Glu + 2Phe + 4Gly + 3Ile + 6Leu + Met
	GUC	Phe, Leu, Ile	Ala, Asp, Gly	Val, Val, Val	4/24 + 2/24 + 2/24 + 2/24 + 4/24 + 3/24 + 6/24 + 1/24
	GUA	Leu, Leu, Ile	Ala, Glu, Gly	Val, Val, Val	
	GUG	Leu, Leu, Met	Ala, Glu, Gly	Val, Val, Val	
Ser	UCU	Pro, Thr, Ala	Phe, Tyr, Cys	Ser, Ser, Ser	4Ala + 2Cys + 2Phe + 2Leu + 4Pro + 4Thr + Trp + 2Tyr + 3STOP
	UCC	Pro, Thr, Ala	Phe, Tyr, Cys	Ser, Ser, Ser	4/24 + 2/24 + 2/24 + 2/24 + 4/24 + 4/24 + 1/24 + 2/24 + 3/24
	UCA	Pro, Thr, Ala	Leu, STOP, STOP	Ser, Ser, Ser	
	UCG	Pro, Thr, Ala	Leu, STOP, Trp	Ser, Ser, Ser	
Pro	CCU	Ser, Thr, Ala	Leu, His, Arg	Pro, Pro, Pro	4Ala + 2His + 4Leu + 2Gln + 4Arg + 4Ser + 4Thr
	CCC	Ser, Thr, Ala	Leu, His, Arg	Pro, Pro, Pro	4/24 + 2/24 + 4/24 + 2/24 + 4/24 + 4/24 + 4/24
	CCA	Ser, Thr, Ala	Leu, Gln, Arg	Pro, Pro, Pro	
	CCG	Ser, Thr, Ala	Leu, Gln, Arg	Pro, Pro, Pro	
Thr	ACU	Ser, Pro, Ala	Ile, Asn, Ser	Thr, Thr, Thr	4Ala + 2Arg + 2Asn + 3Ile + 2Lys + Met + 4Pro + 6Ser
	ACC	Ser, Pro, Ala	Ile, Asn, Ser	Thr, Thr, Thr	4/24 + 2/24 + 2/24 + 3/24 + 2/24 + 1/24 + 4/24 + 6/24
	ACA	Ser, Pro, Ala	Ile, Lys, Arg	Thr, Thr, Thr	
	ACG	Ser, Pro, Ala	Met, Lys, Arg	Thr, Thr, Thr	
Ala	GCU	Ser, Pro, Thr	Val, Asp, Gly	Ala, Ala, Ala	2Asp + 2Glu + 4Gly + 4Pro + 4Ser + 4Thr + 4Val
	GCC	Ser, Pro, Thr	Val, Asp, Gly	Ala, Ala, Ala	2/24 + 2/24 + 4/24 + 4/24 + 4/24 + 4/24 + 4/24
	GCA	Ser, Pro, Thr	Val, Glu, Gly	Ala, Ala, Ala	
	GCG	Ser, Pro, Thr	Val, Glu, Gly	Ala, Ala, Ala	
Tyr	UAU	His, Asn, Asp	Phe, Ser, Cys	Tyr, STOP, STOP	2Cys + 2Asp + 2Phe + 2His + 2Asn + 2Ser + 4STOP
	UAC	His, Asn, Asp	Phe, Ser, Cys	Tyr, STOP, STOP	2/16 + 2/16 + 2/16 + 2/16 + 2/16 + 2/16 + 4/16
Ochre	UAA	Gln, Lys, Glu	Leu, Ser, STOP	Tyr, Tyr, STOP	2Glu + 2Lys + 2Leu + 2Gln + 2Ser + Trp + 4Tyr
Amber	UAG	Gln, Lys, Glu	Leu, Ser, Trp	Tyr, Tyr, STOP	2/15 + 2/15 + 2/15 + 2/15 + 2/15 + 1/15 + 4/15
His	CAU	Tyr, Asn, Asp	Leu, Pro, Arg	His, Gln, Gln,	2Asp + 2Leu + 2Asn + 2Pro + 4Gln + 2Arg + 2Tyr
	CAC	Tyr, Asn, Asp	Leu, Pro, Arg	His, Gln, Gln	2/16 + 2/16 + 2/16 + 2/16 + 4/16 + 2/16 + 2/16
Gln	CAA	Lys, Glu, STOP	Leu, Pro, Arg	His, His, Gln	2Glu + 4His + 2Lys + 2Leu + 2Pro + 2Arg + 2STOP
	CAG	Lys, Glu, STOP	Leu, Pro, Arg	His, His, Gln	2/16 + 4/16 + 2/16 + 2/16 + 2/16 + 2/16 + 2/16
Asn	AAU	Tyr, His, Asp	Ile, Thr, Ser	Asn, Lys, Lys	2Asp + 2His + 2Ile + 4Lys + 2Ser + 2Thr + 2Tyr
	AAC	Tyr, His, Asp	Ile, Thr, Ser	Asn, Lys, Lys	2/16 + 2/16 + 2/16 + 4/16 + 2/16 + 2/16 + 2/16
Lys	AAA	STOP, Gln, Glu	Ile, Thr, Arg	Asn, Asn, Lys	2Glu + Ile + Met + 4Asn + 2Gln + 2Arg + 2Thr + 2STOP
	AAG	STOP, Gln, Glu	Met, Thr, Arg	Asn, Asn, Lys	2/16 + 1/16 + 1/16 + 4/16 + 2/16 + 2/16 + 2/16 + 2/16
Asp	GAU	Tyr, His, Asn	Val, Ala, Gly	Asp, Glu, Glu	2Ala + 4Glu + 2Gly + 2His + 2Asn + 2Val + 2Tyr
	GAC	Tyr, His, Asn	Val, Ala, Gly	Asp, Glu, Glu	2/16 + 4/16 + 2/16 + 2/16 + 2/16 + 2/16 + 2/16
Glu	GAA	STOP, Gln, Lys	Val, Ala, Gly	Asp, Asp, Glu	2Ala + 4Asp + 2Gly + 2Lys + 2Gln + 2Val + 2STOP
	GAG	STOP, Gln, Lys	Val, Ala, Gly	Asp, Asp, Glu	2/16 + 4/16 + 2/16 + 2/16 + 2/16 + 2/16 + 2/16
Cys	UGU	Arg, Ser, Gly	Phe, Ser, Tyr	Cys, Trp, STOP	2Phe + 2Gly + 2Arg + 4Ser + 2Trp + 2Tyr + 2STOP
	UGC	Arg, Ser, Gly	Phe, Ser, Tyr	Cys, Trp, STOP	2/16 + 2/16 + 2/16 + 4/16 + 2/16 + 2/16 + 2/16

Table 3 (*continued*)

Original amino acid	RNA codon	Mutated amino acids translated from the changed RNA codon			Number of translated amino acids with their translation probability
		First position	Second position	Third position	
Opal	UGA	Arg, Arg, Gly	Leu, Ser, STOP	Cys, Cys, Trp	2Cys + Gly + Leu + 2Arg + Ser + Trp 2/8 + 1/8 + 1/8 + 2/8 + 1/8 + 1/8
Trp	UGG	Arg, Arg, Gly	Leu, Ser, STOP	Cys, Cys, STOP	2Cys + Gly + Leu + 2Arg + Ser + 2STOP 2/9 + 1/9 + 1/9 + 2/9 + 1/9 + 2/9
Arg	CGU	Cys, Ser, Gly	Leu, Pro, His	Arg, Arg, Arg	2Cys + 4Gly + 2His + 4Leu + 4Pro + 2Gln + 2Ser + Trp + STOP 2/22 + 4/22 + 2/22 + 4/22 + 4/22 + 2/22 + 2/22 + 1/22 + 1/22
	CGC	Cys, Ser, Gly	Leu, Pro, His	Arg, Arg, Arg	
	CGA	STOP, Arg, Gly	Leu, Pro, Gln	Arg, Arg, Arg	
	CGG	Trp, Arg, Gly	Leu, Pro, Gln	Arg, Arg, Arg	
Ser	AGU	Cys, Arg, Gly	Ile, Thr, Asn	Ser, Arg, Arg	2Cys + 2Gly + 2Ile + 2Asn + 6Arg + 2Thr 2/16 + 2/16 + 2/16 + 2/16 + 6/16 + 2/16
	AGC	Cys, Arg, Gly	Ile, Thr, Asn	Ser, Arg, Arg	
Arg	AGA	STOP, Arg, Gly	Ile, Thr, Lys	Ser, Ser, Arg	2Gly + Ile + 2Lys + Met + 4Ser + 2Thr + Trp + STOP 2/14 + 1/14 + 2/14 + 1/14 + 4/14 + 2/14 + 1/14 + 1/14
	AGG	Trp, Arg, Gly	Met, Thr, Lys	Ser, Ser, Arg	
Gly	GGU	Cys, Arg, Ser	Val, Ala, Asp	Gly, Gly, Gly	4Ala + 2Cys + 2Asp + 2Glu + 6Arg + 2Ser + 4Val + Trp + STOP 4/24 + 2/24 + 2/24 + 2/24 + 6/24 + 2/24 + 4/24 + 1/24 + 1/24
	GGC	Cys, Arg, Ser	Val, Ala, Asp	Gly, Gly, Gly	
	GGA	STOP, Arg, Arg	Val, Ala, Glu	Gly, Gly, Gly	
	GGG	Trp, Arg, Arg	Val, Ala, Glu	Gly, Gly, Gly	

Ala, alanine; Arg, arginine; Asn, asparagine; Asp, aspartic acid; Cys, cysteine; Gln, glutamine; Glu, glutamic acid; Gly, glycine; His, histidine; Ile, isoleucine; Leu, leucine; Lys, lysine; Met, methionine; Phe, phenylalanine; Pro, proline; Ser, serine; Thr, threonine; Trp, tryptophan; Tyr, tyrosine; Val, valine.

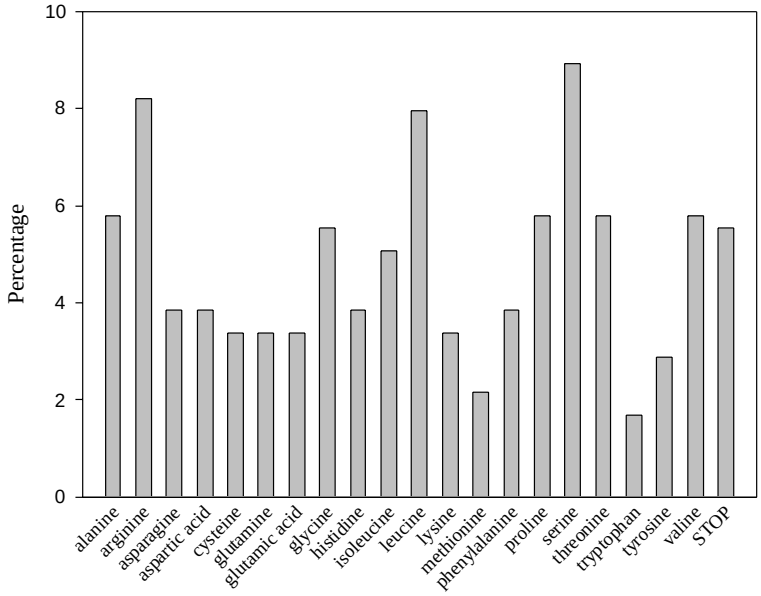

Fig. 1. Theoretical distribution of mutated amino acids for an imaging protein containing equal numbers of 20 types of amino acids.

At this stage, the second approach is more intriguing because a large number of proteins with their mutations have been documented, so a detailed analysis can give us the concepts that may govern the trend of natural mutations. In this study, we apply this approach to the proteins with many mutations, which we have studied in the past using the other computational approach [11–18].

Materials and methods

The following eight proteins with numerous mutations are used in this study. The copper-transporting ATPase 2 (ATP7B, Accession numbers for protein and DNA are P35670 and U11700 [19]) with 125 point mutations [11], the Bruton's tyrosine kinase (BTK, Accession Nos. Q06187 and U10087) with 112 point mutations [12], the haemoglobin α-chain (HBA, Accession Nos. P01922 and J00153) with 130 point mutations [13], the low-density lipoprotein receptor (LDLR, Accession Nos. P01130 and L00352) with 134 point mutations [14], the human p53 protein (p53, Accession Nos. P04637 and M14695) with 192 point mutations [15], the phenylalanine hydroxylase protein (PH4H, Accession Nos. P00439 and K03020) with 187 point mutations [16], the von Hippel–Lindau disease tumor suppressor (VHL, Accession Nos. P40337 and AF010238) with 108 point mutations [17], and the human coronavirus OC43 (OC43, Accession Nos. P36334 and L14643) with 65 point mutations [18].

Translation probability of amino acid from RNA codon. The translation probability is calculated in the same way as shown in Table 2 for all RNA

codons. Totally, there are 183 possible translation probabilities including the STOP codons (Table 3).

Theoretical distribution of mutated amino acids. With the help of probability in Table 3, we can construct a theoretical distribution of mutated amino acids. For example, we imagine that we would have a protein containing equal numbers of 20 types of amino acids. According to the probability in Table 3, we would expect that the mutations would occur in a distribution pattern somewhat similar to that in Fig. 1, say, the mutated amino acids have 5.8% chances of being alanine, and there are 5.5% chances for mutations to result in this imaging protein to be truncated because of the STOP codon. Similarly, with the composition of a protein, we can obtain the theoretical distribution of mutated amino acids such as eight highly mutated proteins in this study.

Actual distribution of mutated amino acids. We can construct the actual distribution of mutated amino acids in proteins with numerous mutations. For instance, the recorded mutations in human p53 protein include the following mutated amino acids, 9 alanines, 9 arginines, 4 asparagines, 9 aspartic acids, 13 cysteines, 10 glutamines, 6 glutamic acids, 11 glycines, 10 histidines, 9 isoleucines, 14 leucines, 7 lysines, 5 methionines, 9 phenylalanines, 14 prolines, 18 serines, 13 threonines, 4 tryptophans, 5 tyrosines, and 9 valines. With these data, we can construct the actual distribution of mutated amino acids and compare it with the theoretical one.

Determination of translation probability of single mutated amino acid at a position. The determination of distribution of translation probability of mutated amino acids from RNA codon is conducted as follows. Taking human p53 protein as an example, the amino acid at position 125 is threonine and its related RNA codon is ACG. A mutation at this position changes threonine to methionine, which has the least probability in Table 2. Thus, this mutation goes along the minimal probability pathway rather than the maximal probability one.

Determination of translation probability of multiple mutated amino acids at a position. There are five mutations occurred at position 245 of human p53 protein, that is, the amino acid glycine is changed into alanine, cys-

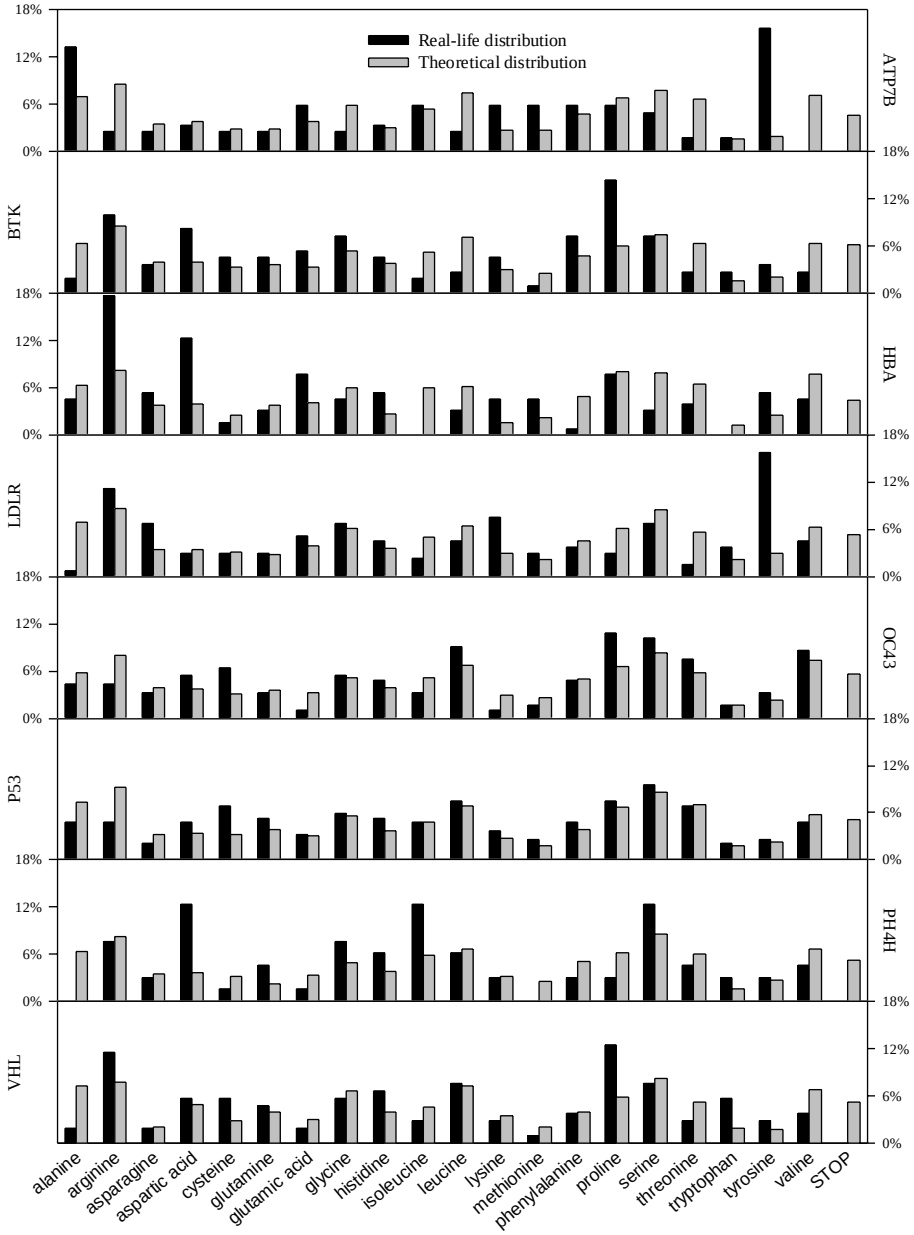

Fig. 2. Theoretical and actual distributions of mutated amino acids for eight highly mutated proteins. ATP7B, copper-transporting ATPase 2; BTK, Bruton's tyrosine kinase; HBA, hemoglobin α chain; LDLR, low-density lipoprotein receptor; p53, human p53 protein; PH4H, phenylalanine hydroxylase protein; VHL, von Hippel–Lindau disease tumor suppressor.

teine, aspartic acid, serine, and valine with the corresponding translation probabilities of 4/24, 2/24, 2/24, 2/24, and 4/24, respectively (the last row in Table 3). In these manners, we can get the distribution of translation probability of mutated amino acids from RNA codons for 192 point mutations in human p53 protein.

Determination of mutated amino acid that cannot be explained by a single change in RNA codon. At the beginning of this study, we could not imagine that there are cases of mutations that cannot be explained from the explicit corresponding relationship between RNA codons and translated amino acids, but we do meet with such cases. For example, the mutation at position 140 in human p53 protein changes threonine to tyrosine, however, we cannot find such a possibility in Table 2. Thus, this is the case, which cannot be explained from the explicit corresponding relationship between RNA codon and translated amino acid.

Statistics. The difference between theoretical and actual translation probabilities are compared using the Wilcoxon signed rank test with SigmaStat for Windows (SPSS, 1992–2003), and the $p < 0.05$ is considered statistically significant.

Results and discussion

Theoretical and actual distributions of mutated amino acids in eight highly mutated proteins

As the evolutionary history is reasonably long, we would expect that the actual distribution of mutated amino acids would approach to the theoretical distribution of mutated amino acids in a protein. Fig. 2 shows the theoretical and actual distributions of mutated amino acids in eight highly mutated proteins. We can see several characteristics in Fig. 2: (i) The theoretical distributions in Fig. 2 are different from those in Fig. 1 and also different one another, this is understandable because the composition of amino acids in each protein differs from that of our imaging protein, which contains equal numbers of 20 types of amino acids, and from other proteins. (ii) There is no actual distribution for STOP codon, which is certainly due to the fact that the premature termination results in a synthesis of deleterious truncated proteins. (iii) Some actual distributions of mutated amino acids are very similar to the theoretical distributions such as isoleucine and threonine in human p53 protein, which at least means that the mutations would not lead to the dysfunction of human p53 proteins if the mutated amino acids would be these types. (iv) Some actual distributions of mutated amino acids are very different from the theoretical distribution such as alanine and arginine in human p53 protein. These phenomena suggest that the mutations lead to the death of human p53 protein if the mutated amino acids would be these types, otherwise we would expect to have seen more records in these mutated amino acids and a smaller difference between the theoretical and actual distributions.

Most missense errors have little effect on protein function, since they only exchange one amino acid for another. Statistical and biochemical studies have revealed non-random patterns in codon assignments. The canonical genetic code is known to be highly efficient in minimizing the effects of mistranslational errors and point mutations, since the biochemical properties of the resulted amino acid are usually very similar to those of the original one when an amino acid is converted to another due to error [20]. Therefore, the implication of the difference between theoretical and actual distributions of mutated amino acids highlights

Table 4
Mutated amino acids that cannot be explained by a single-base change in RNA codons

Protein	Position	Mutation	Change in 2 RNA codes	Change in 3 RNA codes
BTK	594	Glycine → glutamine	GGG → CAG	
OC43	29	Lysine → valine	AAA → GUA	
	173	Glutamine → asparagine	CAA → AAU	
			AAC	
	603	Leucine → threonine	CUU → ACU	
	630	Leucine → threonine	UUA → ACA	
	896	Glutamic acid → cysteine		GAA → UGU
				UGG
	912	Aspartic acid → serine	GAU → UCU	
			AGU	
p53	140	Threonine → tyrosine	ACC → UAC	
	157	Valine → serine	GUC → UCC	
			AGC	
	174	Arginine → histidine		AGG → CAU
				CAC
	247	Asparagine → tryptophan		AAC → UGG
PH4H	157	Arginine → asparagine	AGA → AAU	
			AAC	
VHL	70	Glutamic acid → leucine	GAG → UUG	
			CUG	
	101	Leucine → glycine	CUG → GGG	
	115	Histidine → glutamic acid	CAC → GAA	
			GAG	
	157	Threonine → aspartic acid	ACU → GAU	

BTK, Bruton's tyrosine kinase; OC43, human coronavirus OC43; p53, human p53 protein; PH4H, phenylalanine hydroxylase protein; VHL, von Hippel–Lindau disease tumor suppressor.

the direction of mutations, say, a protein can survive with which type of mutated amino acid.

Mutated amino acids that cannot be explained by single-base change in RNA codon

Table 4 lists the mutated amino acids that cannot be explained by single-base change in the standard genetic codes in the proteins studied herein. Possible explanations for this phenomenon are that the mutated amino acid occurs at the protein level rather than the translation from mRNA, or the mutated amino acid is not related to a single-base change in RNA code, but to two or three (the fourth and fifth columns in Table 4).

Amino-acid misincorporation has been demonstrated during high-level expression [21]. Any errors of translation in the editing-defective cells were due to amino-acid misincorporation, rather than to frameshift errors and an editing deficiency does not contribute to the frequency of spontaneous mutations [22]. Selection at the amino-acid level can influence synonymous codon usage [23]. Non-standard genetic codes are genetic codes in which one or more codons have a different amino-acid assignment from that found in the standard genetic code. The diversity of non-standard genetic codes has found in the modern biosphere. The majority of non-standard codes arise from alterations in the tRNA, with most occurring by post-transcriptional modifications, such as base modification or RNA editing, rather than by substitutions within tRNA anticodons [24]. In some ciliate species, it is found that the UAG and UAA codons encode glutamine, and UGA encodes cysteine and tryptophan [25]. Thus, the

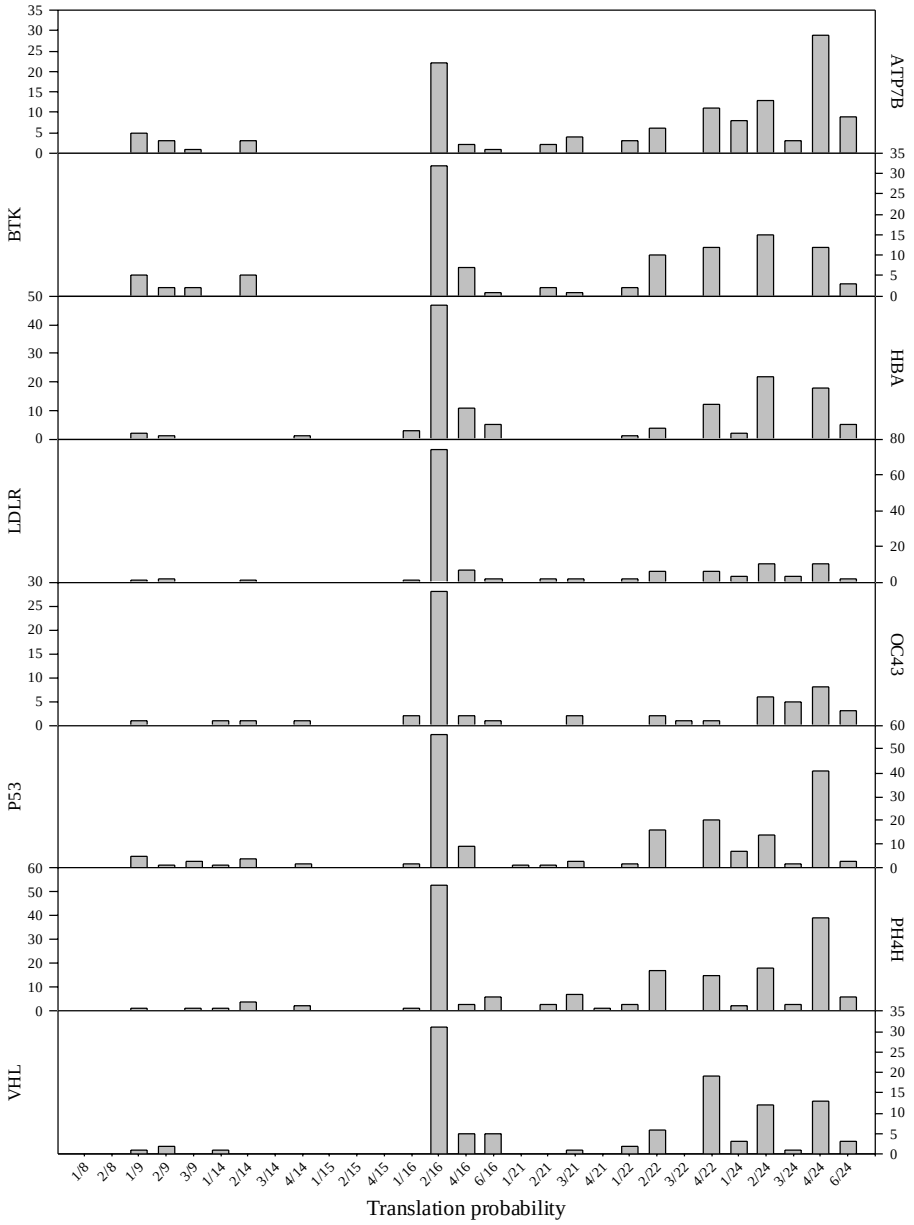

Fig. 3. Translation probability versus frequency of mutated amino acids in eight highly mutated proteins.

findings in Table 4 may involve with the non-standard genetic codes.

Tracing of translation probability from RNA codons to mutated amino acids

Fig. 3 illustrates the translation probability versus the frequency of mutated amino acids in these eight highly mutated proteins. Although these proteins vary regarding their composition of amino acids, their function, their location, and so on, a common pattern can be seen in Fig. 3, which is much clearer than the patterns in Fig. 2, for example, the translation probability of 2/16 has the largest frequency of mutated amino acids. On the other hand, we could expect that the translation probability of 2/16 would be the largest in Fig. 3 because this probability appears most frequently (31.15%) among 183 translation probabilities in Table 3. In terms of percentage of frequency among all the mutations, there is no statistical difference between theoretical and actual situations in Table 5 ($p = 0.109$, Wilcoxon signed rank test). This means that the natural mutation trend goes in principle along the theoretical translation probability listed in Table 3 if the sample is relatively large enough, although we can expect some difference between theoretical and actual mutation frequencies due to the nonsense mutation, dysfunctional mutant, etc.

Table 5
Percentage of mutations in theoretical and actual situations

Translation probability	Theoretical situation based on Table 3		Actual situation in all mutation in Fig. 3	
	Frequency	%	Frequency	%
1/8	4	2.19	0	0.00
2/8	2	1.09	0	0.00
1/9	7	3.83	21	1.99
2/9	4	2.19	11	1.04
3/9	1	0.55	7	0.66
1/14	7	3.83	4	0.38
2/14	5	2.73	18	1.71
3/14	1	0.55	0	0.00
4/14	2	1.09	6	0.57
1/15	1	0.55	0	0.00
2/15	5	2.73	0	0.00
4/15	1	0.55	0	0.00
1/16	2	1.09	9	0.85
2/16	57	31.15	343	32.57
4/16	8	4.37	46	4.37
6/16	2	1.09	21	1.99
1/21	2	1.09	1	0.09
2/21	3	1.64	10	0.95
3/21	3	1.64	20	1.90
4/21	1	0.55	1	0.09
1/22	3	1.64	15	1.42
2/22	7	3.83	67	6.36
3/22	1	0.55	1	0.09
4/22	6	3.28	96	9.12
1/24	5	2.73	25	2.37
2/24	18	9.84	110	10.45
3/24	3	1.64	17	1.61
4/24	19	10.38	170	16.14
6/24	3	1.64	34	3.23
Total	183	100	1053	100

As the genetic code is degenerated for a given polypeptide, a set of synonymous sequences would code the same polypeptide [26]. The relationships between synonymous and non-synonymous substitution rates and between synonymous rate and codon usage bias are important to our understanding of the roles of mutation and selection in the evolutionary process [27]. Synonymous codons differ in their capacity to minimize the effects of errors due to mutation or mistranslation. Natural selection for error minimization at the protein level plays a role in the evolution of coding sequences in *Drosophila* and rodents [28]. At least 10% of variation in codon bias can be explained by mutation pressure [29]. Furthermore, the effect of selection on individual codons changes over time [30]. The selection pressure is for reduced protein synthesis cost, say, most reassignments give amino acids that are less expensive to synthesize. Mitochondrial genetic codes evolve to match the amino-acid requirements of proteins [31].

Our analyses herein point out that the natural mutation trend goes in principle along the theoretical translation probability. In consistent with previous studies, nature should have the intention to spend the least time- and energy-consuming to construct proteins [32].

In conclusion, analyzing the translation probabilities of mutant amino acids governed by the standard genetic codes is performed in the proteins with many mutations. The differences between theoretical and actual distributions of mutated amino acids imply that a protein can survive with which type of mutated amino acids. Some mutated amino acids cannot be explained by single-base errors in the standard genetic codes, which may involve in the non-standard genetic codes. In principle, the natural mutation trend goes along the theoretical translation probability.

References

[1] T.A. Brown, Genome, second ed., BIOS Scientific Publishers, Oxford, 2002.
[2] H. Roy, J. Ling, M. Irnov, M. Ibba, Post-transfer editing in vitro and in vivo by the beta subunit of phenylalanyl-tRNA synthetase, EMBO J. 23 (2004) 4639–4648.
[3] H. Jakubowski, Energy cost of translational proofreading in vivo. The aminoacylation of transfer RNA in Escherichia coli, Ann. N.Y. Acad. Sci. 745 (1994) 4–20.
[4] M. Yarus, S.W. Cline, P. Wier, L. Breeden, R.C. Thompson, Actions of the anticodon arm in translation on the phenotypes of RNA mutants, J. Mol. Biol. 192 (1986) 235–255.
[5] P.N. Allen, H.F. Noller, A single base substitution in 16S ribosomal RNA suppresses streptomycin dependence and increases the frequency of translational errors, Cell 66 (1991) 141–148.
[6] U. von Ahsen, Translational fidelity: error-prone versus hyper-accurate ribosomes, Chem. Biol. 5 (1998) R3–R6.
[7] I. Stansfield, K.M. Jones, P. Herbert, A. Lewendon, W.V. Shaw, M.F. Tuite, Missense translation errors in *Saccharomyces cerevisiae*, J. Mol. Biol. 282 (1998) 13–24.
[8] H. Akashi, A. Eyre-Walker, Translational selection and molecular evolution, Curr. Opin. Genet. Dev. 8 (1998) 688–693.

[9] C.M. Grant, M.F. Tuite, Mistranslation of human phosphoglycerate kinase in yeast in the presence of paromomycin, Curr. Genet. 26 (1994) 95–99.
[10] D.V. Freistroffer, M. Kwiatkowski, R.H. Buckingham, M. Ehrenberg, The accuracy of codon recognition by polypeptide release factors, Proc. Natl. Acad. Sci. USA 97 (2000) 2046–2051.
[11] G. Wu, S. Yan, Determination of amino acid pairs sensitive to variants in human copper-transporting ATPase 2, Biochem. Biophys. Res. Commun. 319 (2004) 27–31.
[12] G. Wu, S. Yan, Determination of amino acid pairs sensitive to variants in human Bruton's tyrosine kinase by means of a random approach, Mol. Simul. 29 (2003) 249–254.
[13] G. Wu, S.M. Yan, Determination of amino acid pairs in human hemoglobulin α-chain sensitive to variants by means of a random approach, Comp. Clin. Pathol. 12 (2003) 21–25.
[14] G. Wu, S. Yan, Determination of amino acid pairs sensitive to variants in human low-density lipoprotein receptor precursor by means of a random approach, J. Biochem. Mol. Biol. Biophys. 6 (2002) 401–406.
[15] G. Wu, S. Yan, Determination of amino acid pairs in human p53 protein sensitive to mutations/variants by means of a random approach, J. Mol. Model. 9 (2003) 337–341.
[16] G. Wu, S.M. Yan, Estimation of amino acid pairs sensitive to variants in human phenylalanine hydroxylase protein by means of a random approach, Peptides 23 (2002) 2085–2090.
[17] G. Wu, S. Yan, Determination of amino acid pairs in Von Hippel–Lindau disease tumour suppressor (G7 protein) sensitive to variants by means of a random approach, J. Appl. Res. 3 (2003) 512–520.
[18] G. Wu, S. Yan, Prediction of amino acid pairs sensitive to mutations in the spike protein from SARS related coronavirus, Peptides 24 (2003) 1837–1845.
[19] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000, Nucleic Acids Res. 28 (2000) 45–48.
[20] H. Goodarzi, H.S. Najafabadi, H.A. Nejad, N. Torabi, The impact of including tRNA content on the optimality of the genetic code, Bull. Math. Biol. (2005), [Available on line July 11].
[21] C.A. Scorer, M.J. Carrier, R.F. Rosenberger, Amino acid misincorporation during high-level expression of mouse epidermal growth factor in *Escherichia coli*, Nucleic Acids Res. 19 (1991) 3511–3516.
[22] J.M. Bacher, V. de Crecy-Lagard, P.R. Schimmel, Inhibited cell growth and protein functional changes from an editing-defective tRNA synthetase, Proc. Natl. Acad. Sci. USA 102 (2005) 1697–1701.
[23] B.R. Morton, Selection at the amino acid level can influence synonymous codon usage: implications for the study of codon adaptation in plastid genes, Genetics 159 (2001) 347–358.
[24] R.D. Knight, S.J. Freeland, L.F. Landweber, Rewiring the keyboard: evolvability of the genetic code, Nat. Rev. Genet. 2 (2001) 49–58.
[25] O.T. Kim, K. Yura, N. Go, T. Harumoto, Newly sequenced eRF1s from ciliates: the diversity of stop codon usage and the molecular surfaces that are important for stop codon interactions, Gene 346 (2005) 277–286.
[26] F. Rodolphe, C. Mathe, Translation conditional models for protein coding sequences, J. Comput. Biol. 7 (2000) 249–260.
[27] K.A. Dunn, J.P. Bielawski, Z. Yang, Substitution rates in *Drosophila* nuclear genes: implications for translational selection, Genetics 157 (2001) 295–305.
[28] M. Archetti, Selection on codon usage for error minimization at the protein level, J. Mol. Evol. 59 (2004) 400–415.
[29] R.M. Kliman, J. Hey, The effects of mutation and natural selection on codon bias in the genes of *Drosophila*, Genetics 137 (1994) 1049–1056.
[30] J.M. Comeron, M. Kreitman, The correlation between synonymous and nonsynonymous substitutions in *Drosophila*: mutation, selection or relaxed constraints? Genetics 150 (1998) 767–775.
[31] J. Swire, O.P. Judson, A. Burt, Mitochondrial genetic codes evolve to match amino acid requirements of proteins, J. Mol. Evol. 60 (2005) 128–139.
[32] G. Wu, S.M. Yan, Randomness in the primary structure of protein: methods and implications, Mol. Biol. Today 3 (2002) 55–69.