Hypothesis

Chain termination codons and polymerase-induced frameshift mutations

Jean-Luc Jestin[a,*], Achim Kempf[b]

[a] *Centre for Protein Engineering and Laboratory of Molecular Biology, Medical Research Council, Hills Road, Cambridge CB2 2QH, UK*
[b] *Corpus Christi College in the University of Cambridge, Cambridge CB2 1RH, UK*

Received 25 September 1997

Abstract The consensus sequence for single-base deletions in non-reiterated runs during in vitro DNA-dependent DNA polymerisation is refined using data available in the literature. This leads to the observation that chain termination codons are hotspots for single-base deletions. The evolutionary implications are discussed in two models which differ in whether polymerases evolved while the genetic code emerged or after the genetic code was fixed. A possible answer to the question 'Why are stop codons just what they are?' is suggested.

Key words: Genetic code; Stop codon; Deletion; Polymerase

1. Assumptions

The mutational spectra of in vitro polymerisation [1–9] for DNA polymerases belonging to families found in at least two of the three living kingdoms [10] are considered here as relevant with respect to a primordial polymerase; we will not take into account DNA polymerases β, as they are family X DNA polymerases so far only found among eukaryotes [11], and HIV reverse transcriptases, which emerged very 'late' in evolution [12].

As it is believed that RNA preceded DNA in evolution [13], data for RNA replicases would be more relevant but are not available; recent evidence shows, however, that DNA and RNA replicases are very closely related [14–16]: a single substitution of a hydroxyl group by a hydrogen atom in the Y639F mutant of T7 RNA polymerase allows a DNA replicase to function as a RNA replicase [17], and a single mutation confers on Moloney murine leukaemia virus reverse transcriptase the ability to replicate RNA [18]; we also note that *Escherichia coli* DNA polymerase I is an accurate RNA-dependent DNA polymerase [19].

2. Polymerase errors

Polymerase-induced mutations are mainly substitutions and frameshifts [1–7]. For the Klenow polymerase domain [1], which has no nuclease domain, as can be assumed for a primordial polymerase, the frameshift error rate is about half the substitution error rate: frameshift mutations therefore represent a significant proportion of the mutations in such systems. Frameshifts result mostly from deletions and additions of one base [1–5]. Crucially, these are highly deleterious by preventing translation in the correct reading frame of the codons downstream of the mutation. Frameshifts occurring in directly repeated and palindromic sequences [8] will be addressed in the discussion. Here, we will focus on frameshifts in non-reiterated runs, where single-base deletions occur far more frequently than single-base additions [1,2,4,7]. Additions will therefore be neglected in the following. For polymerases with and without nuclease domains, the data indicate no significant differences in the consensus sequence for single-base deletions in non-reiterated runs. It has been defined as YR [1], TTR [9], YTG [6] and TR [8]. Using the current data [1–9] we here refine it as YTRV (V = C, A or G; Table 1). These single-base deletions are found to occur generally opposite the purine R (Table 1). The precise mechanism by which these specific sequences alter polymerase fidelity is still unclear [2].

3. The two models

Lessening the phenotypic effects of mutations was suggested to be the major constraint shaping the genetic code [20]. Indeed, the genetic code has been found to reduce significantly the phenotypic effects of transitions: the amino acids that are encoded by two codons have either purines or pyrimidines as third codon bases but not a pyrimidine and a purine [20]. If transitions occur more frequently than transversions, as can be assumed for a primordial polymerase, then the genetic code may be seen as optimised for base substitution tolerance.

Here, we consider the case of polymerase-induced frameshift mutations. The genetic code can be considered as tolerant towards polymerase-induced frameshifts that occur in directly repeated sequences and in palindromic sequences through simple replacement of codons by synonymous codons. It remains to investigate frameshifts occurring in non-reiterated runs. These are mostly single-base deletions occurring in YTRV sequences. If the genetic code is optimised for frameshift tolerance, then it should be possible to code amino acid sequences without using a YTRV sequence, whatever the reading frame.

If the base T is the first base of a codon and in case the previous codon has a pyrimidine as the third codon base, then the amino acid should be encoded without using the six codons TRV; if the base T is the second base of a codon and in case the first base of the following codon is C, A or G, then the amino acid should be encoded without using the codons YTR; if the base T is the third base of a codon and in case the following amino acid has a RVN-type codon, then the amino acid should be encoded without using the eight codons NYT. In summary, TRV, YTR and NYT are potential deletion site codons. Furthermore, their reverse-complementary sequences are also expected to yield deletions during replication.

*Corresponding author. Fax: (44) (1223) 402 140.
E-mail: jlj@mrc-lmb.cam.ac.uk

PII S0014-5793(97)01422-1

Table 1
Sequence contexts of single-base deletion sites in non-reiterated runs obtained in in vitro polymerisation assays for family A and B DNA polymerases

Polymerase	Number of single-base deletions (TR)/number of single-base deletions[a] (number of deletions[b])	Sequence contexts of the deletions occurring in TR sequences[c] (number of deletion occurrences at the site)		
Exonuclease-deficient *E. coli* DNAP II [7]	0.48 (33)	CTGG (4)	TTAC (2)	ATGA (1)
		GTGA (4)	TTGC (2)	ATAG (1)
		ATGT (2)		
DNAP α from KB cells [5]	0.46 (26)	CTGG (7)	TTAC (1)	ATGT (1)
		ATGA (2)	TTAA (1)	
DNAP α from calf thymus [5]	0.54 (13)	GTGA (2)	TTGC (1)	ATGT (1)
		TTAA (1)	CTGG (1)	ATGA (1)
DNAP α from chick embryo [5]	0.56 (16)	CTGG (5)	ATGA (2)	ATGT (2)
Klenow fragment [9]	0.95 (102)	TTGG (42)	TTAC (9)	GTGT (1)
		TTGA (27)	GTAA (2)	TTGT (1)
		CTGC (13)	CTGG (2)	
Klenow fragment [6]	0.76 (193)	CTGC (38)	CTGA (11)	TTAC (1)
		TTGG (34)	ATGA (4)	GTAT (1)
		TTGA (33)	GTAA (4)	GTGT (1)
		CTGG (16)	TTAA (3)	TTAC (1)
Klenow fragment [8]	[d]	TTTA (7)	TTAA (4)	TTCA (2)
		TTGA (5)		
Klenow fragment [3]	0.85 (34)	CTGG (11)	TTGC (4)	CTGG (3)
		GTAA (7)	TTAC (3)	TTAG (1)
Exonuclease-deficient Klenow fragment [1]	0.90 (10)	TTAC (4)	TTAG (1)	CTGG (1)
		TTAA (2)	TTGC (1)	
E. coli DNAP I [1]	0.71 (68)	TTGG (14)	CTGG (4)	TTAC (3)
		TTGA (9)	TTAA (4)	GTGG (1)
		CTGC (8)	GTAA (4)	ATGA (1)
Polymerase domain of *E. coli* DNAP I [1]	0.59 (17)	TTAA (4)	GTAA (2)	ATGA (1)
		CTGG (2)	TTAG (1)	
Thumb mutant of Klenow fragment [3]	0.75 (4)	TTAA (1)	TTGC (1)	GTAA (1)
Exonuclease-deficient T7 DNAP[e] [4]	0.72 (25)	CTGG (7)	TTGC (2)	ATGA (1)
		TTAC (4)	ATAG (1)	
		TTAG (2)	ATGT (1)	

[a]Number of single-base deletions in non-reiterated runs occurring opposite the purine in TR sequences divided by the number of single-base deletions in non-reiterated runs.
[b]Number of single-base deletions in non-reiterated runs studied.
[c]If the deletion occurred opposite a homopolymeric dinucleotide, the deletion site has not been defined. The tetranucleotide sequences given in this table assume that the deletion occurred opposite the 5′ purine of the template homopolymeric dinucleotide [6,9]; this is consistent with TR having been defined as the consensus sequence for single-base deletion sites opposite the purine [8,9].
[d]The quadruplets have been studied in the same sequence context.
[e]In the presence of thioredoxin.

As shown in Fig. 1, the most deleterious codons are then TAA and TAG and their reverse-complementary sequences TTA and CTA that are both potential deletion site codons and reverse-complementary potential deletion site codons. Deletions at codons encoding amino acids are likely to yield non-functional proteins, as all downstream codons are not translated. However, deletions at chain termination codons result at most in the addition of peptides to the proteins' carboxy-termini, thereby likely providing functional proteins (see Scheme 1). Therefore, *by assigning the most deletion-prone sequences to chain termination signals and not to amino acids, the genetic code maximises its frameshift tolerance.*

Further, the fact that the codons TTA and CTA encode leucine, which has the highest, six-fold degeneracy, suggests that frameshift tolerance may be one of the constraints which imposed a high degeneracy on this amino acid.

The codon TGA encodes a chain termination signal as well as the amino acid selenocysteine in eubacteria, archaebacteria and eukaryotes [21,22], except in specific lineages or species [23,24]. This is consistent with our analysis where TGA is found to be a potential deletion site codon but not a reverse-complementary potential deletion site codon. To summarise, the avoidance within coding regions of the deletion-prone YTRV sequences and their tolerance at the ends of coding regions provides a possible answer to the question 'Why are stop codons just what they are?' (see Scheme 1). This explanation indicates that those single-base deletions were sufficiently deleterious within a gene that an emerging code assigning the corresponding codons to the gene's last codon had a selective advantage over other codes, which did not have chain termination codons or which assigned chain termination codons to other codons. On the other hand, it has long been known that base substitutions which yield stop codons are highly deleterious, as they yield truncated proteins [25,26]. Therefore, the codon assignment of chain termination signals may be considered as the balanced result between an optimisation for base substitution tolerance, which minimises the number of stop codons, and an optimisation for frameshift tolerance, which minimises the deleterious effects of single-base deletions.

Our reasoning so far assumed that the genetic code and polymerases have coevolved.

Evidence for the coevolution of the genetic code and amino acid biosynthesis pathways was provided by Wong [27]. Also, Woese, on the basis of experimental data, strongly supported the view that a primitive translation apparatus and the genetic

TTT Phe	**TCT** Ser	TAT Tyr	TGT Cys
TTC	TCC	**TAC**	**TGC**
TTA Leu	TCA	**TAA** CTS	**TGA** CTS & Sec
TTG	TCG	**TAG**	**TGG** Trp
CTT Leu	**CCT** Pro	CAT His	CGT Arg
CTC	CCC	CAC	CGC
CTA	CCA	CAA Gln	CGA
CTG	CCG	CAG	CGG
ATT Ile	**ACT** Thr	AAT Asn	AGT Ser
ATC	ACC	AAC	AGC
ATA	ACA	AAA Lys	AGA Arg
ATG Met	ACG	AAG	AGG
GTT Val	**GCT** Ala	GAT Asp	GGT Gly
GTC	GCC	GAC	GGC
GTA	GCA	GAA Glu	GGA
GTG	GCG	GAG	GGG

Fig. 1. A representation of the genetic code highlighting potential deletion site codons. The potential deletion site codons NYT, YTR and TRV (see text) are noted in bold and their reverse-complementary sequences are underlined. TAA and TAG as well as their reverse-complementary sequences TTA and CTA are both potential deletion site codons and reverse-complementary potential deletion site codons, i.e. hotspots for single-base deletions in non-reiterated runs. CTS: chain termination signal.

code have coevolved [28,29]. More generally, the emergence of the genetic code may be seen as a key step in the evolution of self-reproductive systems by allowing a cooperation between nucleic acids and proteins; this cooperation is highlighted by RNA-directed protein synthesis and by polymerase-directed nucleic acid replication [30]. Furthermore, Epstein suggested that the genetic code evolved to reduce the effects of mutations arising during replication [31]. The hypothesis that the genetic code may have coevolved with a primordial polymerase before the genetic code was fixed does therefore appear as reasonable.

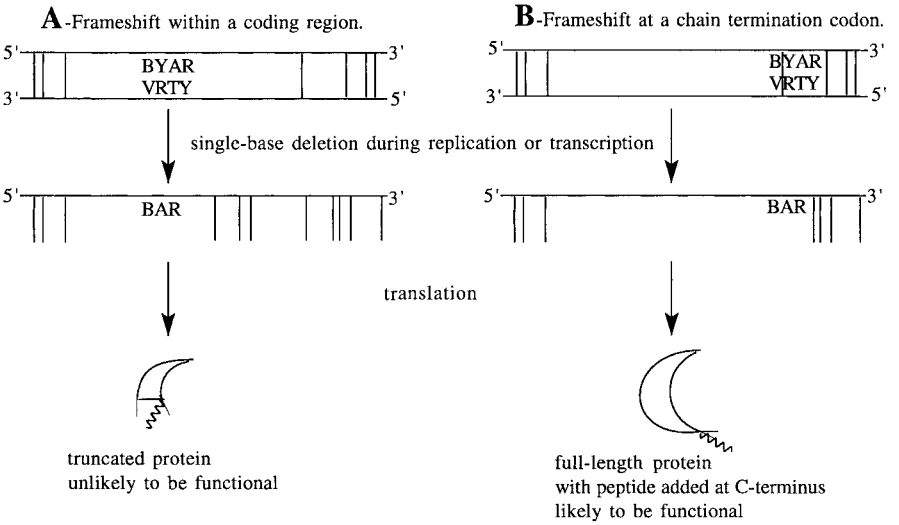
However, the alternative model must also be considered, namely, that protein polymerases evolved after the fixation of the genetic code. In this case, the genetic code evolved in the presence of ribozymes as polymerases. *The evolutionary pressure on protein polymerase fidelity would then be weaker at chain termination codons than at codons that encode amino acids* – again because single-base deletions at stop codons are likely to yield fully functional proteins. This constitutes an alternative interpretation for our observation that chain termination codons are hotspots for single-base deletions.

4. Discussion

The model according to which the codon assignment of chain termination signals optimises frameshift mutation tolerance provides new support to the theory stating that the genetic code has been selected so as to minimise the effects of errors [20,25,28,29,31].

This is consistent with the theory of the 'frozen accident', which states that "the code is universal because at the present time any change would be lethal, or at least very strongly selected against" [32]. Although evidence for an incomplete fixation of the genetic code has been provided [33], the various known genetic codes present only minor differences and the general shape of the universal genetic code is not altered. To this extent, Crick's statement still remains an excellent hypothesis.

Further, according to Crick, "there is no reason to believe, however, that the present code is the best possible [...]. Instead, it may be frozen at a local minimum which it has reached by a rather random path" [32]. "[This] theory seems plausible but as a theory it suffers from a major defect: it is too accommodating. In a loose sort of way it can explain anything" [32]. Evidence for a primitive 'operational' code from which the universal genetic code may derive has been provided [34,35]. Direct interactions between ribonucleic acids and amino acids



Scheme 1. Most single-base deletions are less deleterious at chain termination codons than at codons encoding amino acids. The observation that chain termination codons are hotspots for single-base deletion (see Fig. 1) can then be interpreted in two ways. Assuming that polymerases coevolved with the genetic code, the codon assignment of chain termination signals is seen to minimise the deleterious effects of polymerase-induced frameshift mutations. Alternatively, assuming that polymerases evolved after fixation of the genetic code, the observation indicates that the selection pressure on polymerase fidelity was weaker at chain termination codons than at codons encoding amino acids. The horizontal bars represent nucleic acid strands, the vertical ones chain termination codons, the zigzag patterns peptides. YTRV is the consensus sequence for single-base deletions in non-reiterated runs, the deletion occurring opposite the purine R (see text and Table 1); BYAR is its reverse-complementary sequence. V = C,A,G and B = T,C,G.

are involved and may represent the basis for a 'stereochemical theory'. Wong's theory where it is convincingly argued that biosynthetically related amino acids have closely related codons also sheds some light on the way the genetic code has been attained [27,36].

To conclude, within the model of the genetic code coevolving with the replication or translation machineries, we suggest here that theories of mutation effect minimisation based on experimental data provide a rationale for the way the genetic code has reached the minimum or a local minimum. However, within the alternative model of protein polymerase and ribosome evolving after the genetic code was fixed, the same experimental data on errors occurring during replication and translation have a different interpretation: leaving open the question of why the codon assignments are just what they are, the data then indicate a sequence-dependent modulation of the selection pressure on replication and translation fidelities due to the shape of the genetic code.

Acknowledgements: We are grateful to an anonymous referee for his very useful criticisms. We are also indebted to Greg Winter, Henri Buc, Andrew Travers, Peter Wang, Pierre Legrain and Aaron Klug for critical comments on the manuscript and thank Adrian Kent and Eric De La Fortelle for advice. J.L.J. was supported by an EMBO fellowship.

References

[1] Bebenek, K., Joyce, C., Fitzgerald, M. and Kunkel, T. (1990) J. Biol. Chem. 265, 13878–13887.
[2] Bell, J., Eckert, K., Joyce, C. and Kunkel, T. (1997) J. Biol. Chem. 272, 7345–7351.
[3] Minnick, D., Astatke, M., Joyce, C. and Kunkel, T. (1996) J. Biol. Chem. 271, 24954–24961.
[4] Kunkel, T., Patel, S. and Johnson, K. (1994) Proc. Natl. Acad. Sci. USA 91, 6830–6834.
[5] Kunkel, T. (1985) J. Biol. Chem. 260, 12866–12874.
[6] Papanicolaou, C. and Ripley, L. (1989) J. Mol. Biol. 207, 335–353.
[7] Cai, H., Yu, H., McEntee, K., Kunkel, T. and Goodman, M. (1995) J. Biol. Chem. 270, 15327–15335.
[8] Wang, F. and Ripley, L. (1994) Genetics 136, 709–719.
[9] De Boer, J. and Ripley, L. (1988) Genetics 118, 181–191.
[10] Braithwaite, D. and Ito, J. (1993) Nucleic Acids Res. 21, 787–802.
[11] Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. and Sonnhammer, E. (1992) Protein Sci. 1, 1677–1690.
[12] Eigen, M. and Nieselt-Struwe, K. (1990) AIDS 4, S85–S93.
[13] Darnell, J. and Doolittle, W. (1986) Proc. Natl. Acad. Sci. USA 83, 1271–1275.
[14] Delarue, M., Poch, O., Tordo, N., Moras, D. and Argos, P. (1990) Protein Eng. 5, 461–467.
[15] Joyce, C. and Steitz, T. (1994) Annu. Rev. Biochem. 63, 777–822.
[16] Joyce, C. (1997) Proc. Natl. Acad. Sci. USA 94, 1619–1622.
[17] Sousa, R. and Padilla, R. (1995) EMBO J. 14, 4609–4621.
[18] Gao, G., Orlova, M., Georgiadis, M., Hendrickson, W. and Goff, S. (1997) Proc. Natl. Acad. Sci. USA 94, 407–411.
[19] Ricchetti, M. and Buc, H. (1993) EMBO J. 12, 387–396.
[20] Goldberg, A. and Wittes, R. (1966) Science 153, 420–424.
[21] Böck, A., Forchhammer, K., Heider, J., Leinfelder, W., Sawers, G., Veprek, B. and Zinoni, F. (1991) Mol. Microbiol. 5, 515–520.
[22] Hatfield, D. and Diamond, A. (1993) Trends Genet. 9, 69–70.
[23] Osawa, S., Muto, A., Ohama, T., Andachi, Y., Tanaka, R. and Yamao, F. (1990) Experientia 46, 1097–1106.
[24] Jukes, T. and Osawa, S. (1990) Experientia 46, 1117–1126.
[25] Sonneborn, T. (1965) in: Evolving Genes and Proteins (Bryson, V. and Vogel, H.J., Eds.), pp. 377–397, Academic Press, New York.
[26] Shcherbak, V.I. (1989) J. Theor. Biol. 139, 283–286.
[27] Wong, J. (1975) Proc. Natl. Acad. Sci. USA 72, 1909–1912.
[28] Woese, C. (1965) Proc. Natl. Acad. Sci. USA 54, 1546–1552.
[29] Woese, C. (1973) Naturwissenschaften 60, 447–459.
[30] Eigen, M. (1971) Naturwissenschaften 1971, 465–523.
[31] Epstein, C.J. (1966) Nature 210, 25–28.
[32] Crick, F.H.C. (1968) J. Mol. Biol. 38, 367–379.
[33] Osawa, S., Jukes, T.H., Watanabe, K. and Muto, A. (1992) Microbiol. Rev. 56, 229–264.
[34] Hou, Y.M. and Schimmel, P. (1988) Nature 333, 140–145.
[35] Rodin, S.N. and Ohno, S. (1997) Proc. Natl. Acad. Sci. USA 94, 5183–5188.
[36] Di Giulio, M. (1997) Trends Biochem. Sci. 22, 49.