 doi:10.20944/preprints202004.0533.v1

Optimal evolution of the standard genetic code

Michael Yarus

Department of Molecular, Cellular and Developmental Biology

University of Colorado Boulder

Boulder, CO 80309-0347

Tel: 303 492-8376
Mobile: 303 817-6018
Email: yarus@colorado.edu
Keywords: coding, codon, triplet, evolve, wobble

Abstract

The Standard Genetic Code (SGC) exists in every organism known on Earth. SGC evolution via early unique codon assignment, then later wobble, yields coding resembling the near-universal code. Below, later wobble also creates an optimal route to accurate codon assignment. This assignment time matches a previous mean time for ordered codes, exhibiting ≥ 90% of SGC order. Accurate evolution is also accessible, sufficiently frequent to appear in populations of 10^3 to 10^4 codes. SGC-like coding capacity, code order and assignments therefore arise together, in one attainable evolutionary intermediate. Examples, which plausibly resemble coding at evolutionary domain separation, are characterized.

Introduction

Early evolution of the Standard Genetic Code (SGC) has been computed (Yarus 2020) by dividing code formation into time slices (passages). During a passage, with specified probability, coding triplets may either be assigned, capture mutationally related triplets for their preexisting assignment or can decay, losing assigned meaning. But nothing may happen during a passage. This procedure yields normal dynamic phenomena, like first and second order rates, as well as standard near-steady states. It is mathematically equivalent to defining typical first and second order kinetic constants for initiation, decay and capture of codon assignments (Yarus 2020).

SGC-like coding tables arise by combining stereochemical initial codon assignment (Yarus 2017, 10% randomness allowed) with coevolutionary capture (Wong 1981) that prefers amino acids with similar polar requirements (Woese et al. 1966; Mathew and Luthey-Schulten 2008). More particularly, in order to fill a coding table, wobble must arise late; for example, appearing after 20 amino acids are assigned. This leaves modern initiation and termination for a later origin, consistent with their unconserved, and thus late-arising, molecular components (Yarus 2020). Below, late wobble not only provides access, but also an accessible, optimal route to SGC-like order with amino acid coding capacity.

Simple wobble is used. To suit primordial coding, only natural unmodified nucleotides are assumed. Thus, wobble implies only that U:G and G:U pairs are allowed at third codon nucleotide positions, as Crick first proposed (Crick 1966). Thus XY***U*** and XY***G*** may be read either by normal base-pairing (***A*** and ***C***, respectively) or by wobble pairings (***G*** and ***U***, respectively).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 April 2020 doi:10.20944/preprints202004.0533.v1

Ingredients for an optimum. In Fig. 1A, late-wobbling coding table evolution is shown in a fashion designed to clarify approach to an SGC-like coding table.

The fraction of coding tables encoding 20 functions or more rise with time (≥ 20 fn; on the leftward ordinate), after a lag to accumulate enough assigned triplets for 20 distinct assignments.

The fraction of coding tables correctly assigned are about an order smaller and descend with time (on the rightward ordinate), after the first near-complete tables appear around 60 passages. Eventually, Fig. 1A assignments that are 10% at random (rather than wholly SGC-like) will be made in error.

Moreover, the SGC has many wobble assignments. But an early non-wobble coding table, evolving at 60 passages, is already 67% full. Thereafter, decreasing room limits added wobble. Thus, as the table is filled before wobble is instituted, wobble declines and resemblance to the SGC itself ultimately suffers.

Three descending curves show coding accuracy: 2 or fewer assignment errors (≤ 2 mis; top line), 1 or fewer misassignments (≤ 1 mis; middle) or fully SGC-like coding with no assignment error (0 mis; bottom). In the latter case, requiring complete assignment precision, we touch near the SGC itself.

Thus, Fig. 1A shows a crucial conflict. Sufficient capacity for realistic coding increases with time; accuracy decreases with time. So, there is an optimum, when accuracy and coding capacity co-exist.

The optimum defined. Fig. 1B combines these coding criteria. The rising plot of capacity for ≥ 20 encoded functions is repeated to help time other events.

The other three lines, with maxima, show the fraction of codes that have both capacity to encode ≥ 20 functions and ≤ 2 assignment errors (top), ≥ 20 functions and ≤ 1 error (middle) and ≥ 20 functions with complete SGC-like codon assignment (bottom). Notably, most probable times for both qualities are similar: 100 passages (≤ 2 mis), 100-120 passages (≤1 mis) and 120 passages (0 mis). Optimal durations allow 39 to 44 initial triplet assignments, under these conditions (Fig. 1A, legend).

Accessibility of SGC-like codes. In Fig. 1B, the abundance of competent codes is determined. This implies a code population size that must be explored to find SGC-like coding (Fig. 1C).

P_{obs} = probability of observation in E independent evolutions, with event abundance/probability = P_{event}:

$$(1 - P_{event})^E = (1 - P_{obs})$$

$$E = \frac{\ln(1 - P_{obs})}{\ln(1 - P_{event})} \cong -\frac{\ln(1 - P_{obs})}{P_{event}}$$

where the latter equation is accurate for somewhat rare events, $P_{event} <\approx 0.1$. For even odds of observation, $P_{obs} = 0.5$:

$$E = {\ln 2}/{P_{event}}$$

Abundance of accurately-formed codes, P_{event} (Fig. 1B), implies plausible population size E.

In Fig. 1C, fraction of codes specifying ≥ 20 functions again serves as time reference. Importantly, populations of about 300 (≤ 2 mis) codes, 1200 (≤ 1 mis) codes or of 8700 (0 mis) codes would be required, at minima in Fig. 1C, to find codes that closely resemble the SGC, with even odds. SGC-like

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 April 2020 doi:10.20944/preprints202004.0533.v1

coding can therefore exist in biologically conceivable evolving populations, despite astronomically vast ensembles of possible coding schemes.

These results are consistent with previous searches (Yarus 2020), in which codes with 1 or 2 misassignments were found among 600 coding tables, evolving under similar late wobble conditions. In those prior results, however, SGC-like code order was sought, rather than explicit assignment accuracy.

A four-fold optimum. So, remarkably: there is a time (≈115 passages; Fig. 1B, 1C; Yarus 2020) and a coding state (after 42 initial assignments under these conditions; Fig. 1A) when a nascent coding table, having just adopted wobble, simultaneously possesses near-optimal spacing (identical assignments in related codons), near-optimal chemical order (related triplets associated with similar polar requirements), and SGC-like triplet sequences, the latter extending to total identity: codes with no assigned codons differing from the SGC (Fig. 1B,1C, 2). Such coding tables encode 20 functions (Fig. 1A, 1B, 1C) and for example, might specify all amino acids. Such competent coding precisely overlapping the SGC appears with probability ≈ 8×10^{-5} (Fig. 1B).

Filling the coding table. To entirely resemble the SGC, one also wants full codes; most triplets assigned, but room left for late assignment of modern initiation and termination codons, and perhaps a few others. In shifting to optimal time (Fig. 1B, 1C), unassigned triplets increase: at the 177 passage mean time (Yarus 2020), 5.4 triplets were unassigned, and at the present 115 passage optimum, 8.5 triplets are yet to be accounted for, on average. However, free triplets are also distributed. For the optimal time in Fig. 1C, 0.24 of all evolutions have 0 to 4 triplets free (using assignments for modern initiation and termination to define this hypothetical target). Because coding without mis-assignment has the same free distribution, this implies that codes that also have an appropriate set of unassigned triplets might be 0.24 the number in Fig. 1B, 1C. Using the above abundance equation – nearly full, highly ordered, 20 function, accurately assigned codes with appropriate room for known late functions appear among:

$$\frac{8700}{0.24} = 3.6\ x\ 10^4\ independent, late\ wobbling\ evolutions.$$

If there must also be room for other late functions assigned by different means, the code population required for even odds would contain between 8.7×10^3 and 3.6×10^4 independent codes. Notably, the individuals required to present codes will be greater than the number of independent codes. But this is a feasible biological ensemble; in fact, perhaps a small group, if code-bearers are microorganisms.

A coding example. To make numerical results tangible, consider an explicit example. Fig. 2 is an evolved coding table from the likely range above. It has been colored to display amino acid chemistry as polar requirement (Woese et al. 1966; Mathew and Luthey-Schulten 2008): blue for the most hydrophobic, through light blue, then gray, beige, orange and red, the latter for the most hydrophilic amino acids. Parenthetical numbers beside amino acid names are corrected polar requirements (Mathew and Luthey-Schulten 2008). Fig. 2's encoding was the 12,804th in a series of 115 passage evolutions, has 6 unassigned triplets, encodes 20 functions, agrees completely with SGC assignments, and has SGC-like order: compactness (spacing = 0.947; where random coding = 0.0 and SGC = 1.0), SGC-like distance (distance = 0.962), and SGC-like chemical order (dPR = 0.963).

Fig. 2 visually confirms that code order and assignment accuracy coexist at 115 passages. In its calculation, 26 such inerrant coding tables were evolved. These were found among ≈ 9000 evolutions with probability 0.5, agreeing with Fig. 1B's 8700. This implies an abundance of 7.7×10^{-5}, again agreeing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 April 2020 doi:10.20944/preprints202004.0533.v1

with Fig. 1B's 8 x 10^{-5}. Fig. 2 varies from the SGC with gaps in canonical Pro and Ala boxes, incomplete encoding at one termination (Ter) codon and a complex, possibly late-evolving amino acid (Trp; Grosjean and Westhof 2016). Except for such arguably realistic incompletions, Fig. 2 is SGC-like.

Time for SGC-like codes. Starting bloc selection (Yarus 2018) identifies early biological development as the ideal time for selection of a desired improvement. Were resemblance to the SGC selectable, such selection would work best early in coding table evolution (Fig. 1B, 1C). Accordingly, the SGC may be a new instance of starting bloc selection.

Evolution under these conditions (Fig. 1A legend) can be summarized, spanning a possible starting bloc. Early events fall readily into one of three approximately equal eras. During era 1 (ends at gray arrow 1, Fig. 1C), partial coding tables are filled to produce mature coding capacity (Fig. 1A). Era 1 codes likely compete on the basis of coding capacity. However, at 60 passages and 28 initial assignments, 20-function codes still comprise < 1% of the population.

During the second, optimal era (end at gray arrow 2, Fig. 1C), passage of about the same amount of additional time, and 13 more initial triplet assignments produce an optimum. Twenty-function late wobble coding sharply increases, 20- to 30-fold (Fig. 1B). Such coding can be highly ordered (Fig. 2): with identical assignments grouped, chemically similar amino acids associated with related triplets, and distance to the SGC short. A particularly interesting short distance exists in 7.7 x 10^{-5} of coding tables matching SGC capacity, order and assignments simultaneously (gray arrow 2, Fig. 1C). Very SGC-like codes could be selected among 10^4 independent codes or a few-fold more. However, if one or two differing codon assignments are tolerable, hundreds or thousands of codes could be sufficient (gray arrow 2, Fig. 1C).

Fig. 2 shows an evolved code example with accurate SGC-like assignments, from the end of era 2. The domains of life use similar initiation and termination triplets, but different mechanisms and molecules for interpreting them (Yarus 2020). Thus, coding for translation initiation and termination was defined before mature initiation and termination mechanisms were settled. Fig. 2 therefore may parallel the genetic code near domain separation, when bacterial, archeal and eukaryotic domains diverged.

Third-era, averaging evolution, ends with 50 mean initial assignments and 20 mean era 2-encoded functions (Yarus 2020), again, after passage of another, similar, era (at gray arrow 3, Fig. 1C). It seems likely that the SGC was completed in this era, adding definitive 21st and 22nd functions, initiation and termination. But fully competent, fully ordered, fully accurate codes are ≈ 4-fold rarer at gray arrow 3 than arrow 2, defining a past optimum.

Given a credible duration for any one event, this reasoning will estimate real times on an early Earth. It does not seem overly optimistic to suppose that this will be achieved.

Figure Legends.

Figure 1A. Fraction of coding tables with coding capacity (≥ 20 encoded functions: left ordinate) or SGC-like assignments (0 mis, no misassignment; ≤ 1 mis, ≤ 1 misassignment; ≤ 2 mis, ≤ 2 misassignments: right ordinate), versus evolutionary time as passages. 100,000 late wobble evolutions via Coevo_PR captures: coevolution (Wong 1981; Amirnovin 1997; Ronneberg et al. 2000) with chemical similarity (Freeland and Hurst 1998; Massey 2019) to direct amino acid choice. Probabilities per passage are Pmut

= 0.04 for mutational capture, Pdecay = 0.04 for assignment loss, Pinit = 0.6 for initial triplet assignment, Prand = 0.1 for random assignment (Yarus 2020).

Figure 1B. Fraction of coding tables with coding capacity (≥ 20 encoded functions: left ordinate) or with both coding capacity and accurate assignment, versus evolutionary time as passages (labeling, history and parameters as in Figure 1A). "order" indicates mean and sem for ≥ 20 functions with joint code order ≥ 0.9 of the SGC; "mean" indicates mean and sem for ≥ 20 encoded functions (Yarus 2020).

Figure 1C. Fraction of coding tables with coding capacity (≥ 20 encoded functions: left ordinate) or number of independent codes examined to find a code with indicated accuracy half the time (right, logarithmic ordinate), versus evolutionary time as passages (labeling, history and parameters as in Figure 1A, 1B, save for numbered gray arrows at bottom, which mark their leftward eras for discussion).

Figure 2. Evolved coding example from optimal time, with 20 encoded functions and no misassignments. Evolved for 115 passages using history and parameters as in Fig. 1A. More detail is in the text.

References.

Amirnovin R. 1997. An Analysis of the Metabolic Theory of the Origin of the Genetic Code. *J Mol Evol* **44**: 473–476.

Crick FH. 1966. Codon-anticodon pairing: the wobble hypothesis. *J Mol Biol* **19**: 548–555.

Freeland SJ, Hurst LD. 1998. The genetic code is one in a million. *J Mol Evol* **47**: 238–248.

Grosjean H, Westhof E. 2016. An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Res* **44**: 8020–8040.

Massey SE. 2019. Genetic Code Error Minimization as a Non-Adaptive But Beneficial Trait. *J Mol Evol* **87**: 4–6.

Mathew DC, Luthey-Schulten Z. 2008. On the physical basis of the amino acid polar requirement. *J Mol Evol* **66**: 519–528.

Ronneberg TA, Landweber LF, Freeland SJ. 2000. Testing a biosynthetic theory of the genetic code: fact or artifact? *Proc Natl Acad Sci U A* **97**: 13690–5.

Woese CR, Dugre DH, Saxinger WC, Dugre SA. 1966. The molecular basis for the genetic code. *Proc Natl Acad Sci U A* **55**: 966–974.

Wong JT-F. 1981. Coevolution of genetic code and amino acid biosynthesis. *Trends Biochem Sci* **6**: 33–36.

Yarus M. 2018. Eighty routes to a ribonucleotide world; dispersion and stringency in the decisive selection. *RNA N Y N* **24**: 1041–1055.

Yarus M. 2020. Evolution of the standard genetic code. *bioRxiv* 2020.02.20.958546.

Yarus M. 2017. The Genetic Code and RNA-Amino Acid Affinities. *Life* **7**: 13.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 April 2020 doi:10.20944/preprints202004.0533.v1

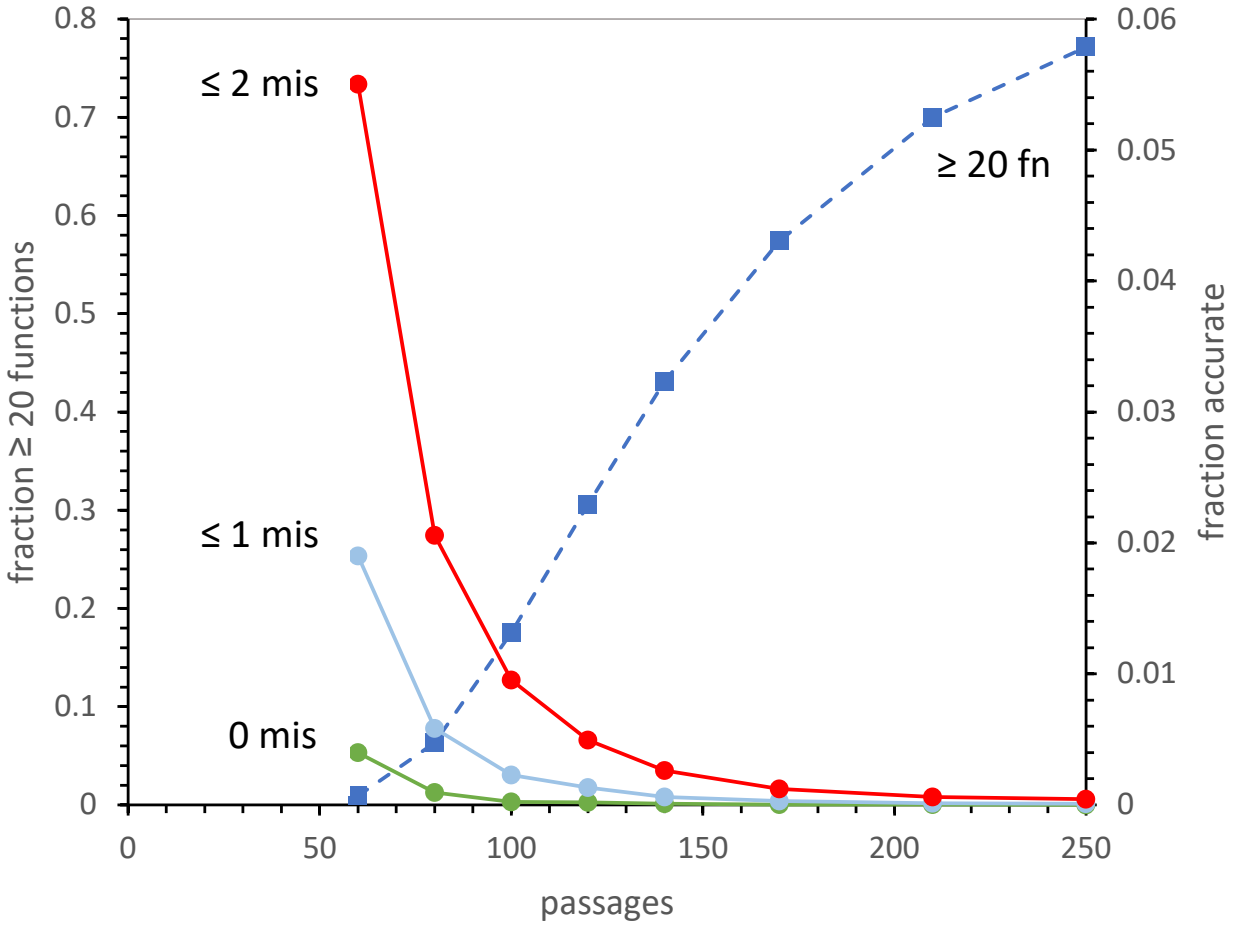

Figure 1A

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 April 2020 doi:10.20944/preprints202004.0533.v1

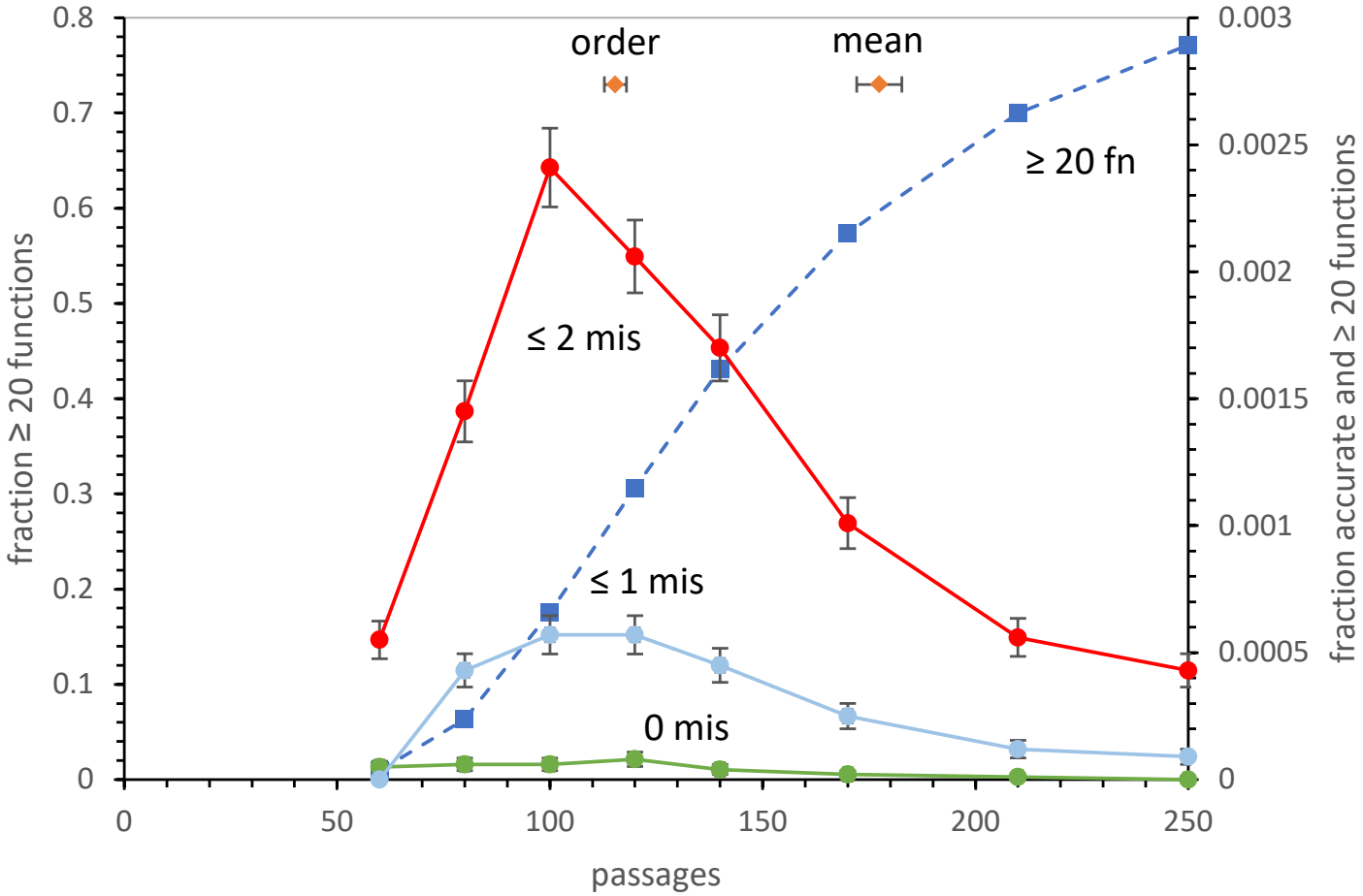

Figure 1B

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 April 2020 doi:10.20944/preprints202004.0533.v1

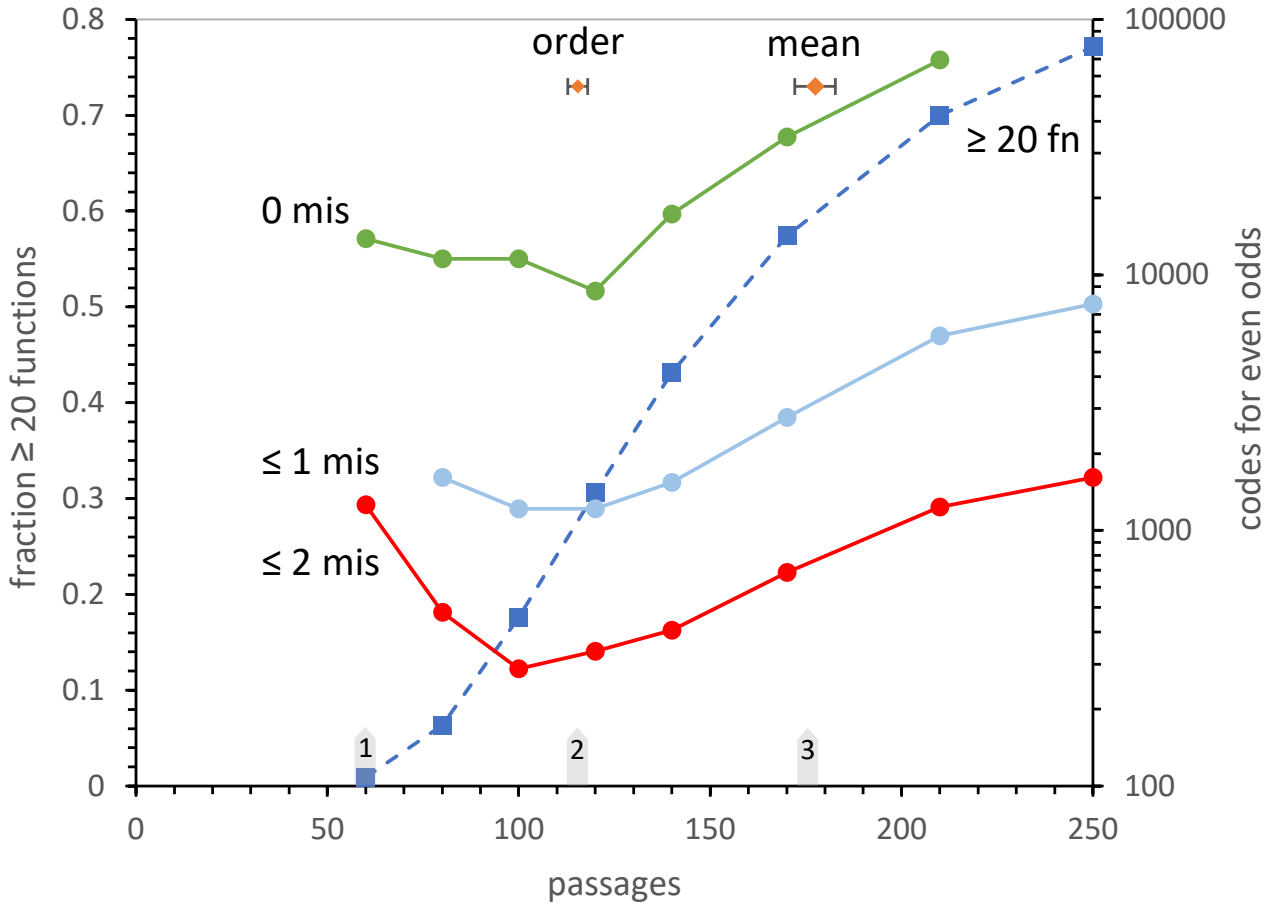

Figure 1C

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 April 2020 doi:10.20944/preprints202004.0533.v1

UUU	Phe (4.5)	UCU	Ser (7.5)	UAU	Tyr (7.7)	UGU	Cys (4.3)
UUC	Phe (4.5)	UCC	Ser (7.5)	UAC	Tyr (7.7)	UGC	Cys (4.3)
UUA	Leu (4.4)	UCA	Ser (7.5)	UAA	Ter	UGA	--
UUG	Leu (4.4)	UCG	Ser (7.5)	UAG	Ter	UGG	--
CUU	Leu (4.4)	CCU	Pro (6.1)	CAU	His (7.9)	CGU	Arg (8.6)
CUC	Leu (4.4)	CCC	Pro (6.1)	CAC	His (7.9)	CGC	Arg (8.6)
CUA	Leu (4.4)	CCA	--	CAA	Gln (8.9)	CGA	Arg (8.6)
CUG	Leu (4.4)	CCG	--	CAG	Gln (8.9)	CGG	Arg (8.6)
AUU	Ile (5.0)	ACU	Thr (6.2)	AAU	Asn (9.6)	AGU	Ser (7.5)
AUC	Ile (5.0)	ACC	Thr (6.2)	AAC	Asn (9.6)	AGC	Ser (7.5)
AUA	Ile (5.0)	ACA	Thr (6.2)	AAA	Lys (10.2)	AGA	Arg (8.6)
AUG	Ini/Met(5.0)	ACG	Thr (6.2)	AAG	Lys (10.2)	AGG	Arg (8.6)
GUU	Val (6.2)	GCU	Ala (6.5)	GAU	Asp (12.2)	GGU	Gly (9.0)
GUC	Val (6.2)	GCC	Ala (6.5)	GAC	Asp (12.2)	GGC	Gly (9.0)
GUA	Val (6.2)	GCA	--	GAA	Glu (13.6)	GGA	Gly (9.0)
GUG	Val (6.2)	GCG	--	GAG	Glu (13.6)	GGG	Gly (9.0)

Figure 2