Original Study • DOI: 10.2478/biocosmos-2023-0001 • BIOCOSMOS • 1 • 2023 • 1-11

BioCosmos

HIGHLY SIMILAR AVERAGE COLLATERAL EFFECT OF SYNONYMOUS MUTATIONS ACROSS ALTERNATIVE READING FRAMES: A POTENTIAL ROLE IN EVOLVABILITY

Stefan Wichmann[1] and Zachary Ardern[1,2]*

[1]Technical University of Munich, Munich, Germany
[2]Wellcome Sanger Institute, Cambridge, United Kingdom

Abstract

Synonymous mutations in a protein coding gene lead to a remarkably similar average "collateral" mutation effect size across alternative reading frames (1). Here we quantify the rarity of this feature among possible block structure codes as 0.77%. Then we develop a simple model of evolutionary search with two types of mutation. Across different mutation step sizes and ratios of the two types, the fitness-maximizing region corresponds to using a single average mutation value. The analogous constant average collateral mutation effect observed for the standard genetic code may likewise facilitate evolutionary search in alternative frame sequences.

Keywords

genetic code optimality • evolvability • gene origins

Introduction

Protein coding in alternative reading frames of known genes is receiving increased attention. This is a return to a topic that was the subject of significant interest early in the development of modern molecular genetics. Overlapping genes (OLGs) were first found in bacteriophages (2), and among biologists they are often assumed to be restricted to viruses. There are many prominent examples in viruses beyond bacteriophages including in the pandemic viruses HIV-1 and SARS-CoV-2 (3–7), but they have also recently been discovered in diverse cellular organisms, including bacteria (8–13), archaea (14), and mammals, including humans (15–19). Research on OLGs within five years of their initial discovery investigated diverse topics, including triple overlaps (20), information theory (21–23), discussion of their evolution (24, 25), and the proposal that they may be widespread throughout life (26). A recent review has collated the evidence for overlapping genes, with discussion of their biological roles and potential biosynthetic applications (27). In addition to the phenomenon of overlapping genes, alternative frame sequences can be incorporated into proteins and thus contribute to protein sequence novelty in other ways, which we will briefly discuss.

Is there a functional reason for the maintenance of overlapping genes in genomes or the evolutionary incorporation of sections of alternative frame sequences into proteins? It is often thought that a need for genome compression provides a selective pressure for overlapping genes. A study directly addressing this question, however, suggests that evolutionary pressure for smaller genomes does not explain OLG distribution in viruses; instead, their role in the evolution of functional novelty may be more important (28). Connected to this, there has been a revolution in our understanding of protein evolution, and now many instances of genuinely "de novo" origin are known (29). Alternative frame sequences have been proposed as a source of evolutionarily novel genes, originating through a process called "overprinting" (30). They were mentioned in a foundational text for the standard hypothesis that most genes arise through duplication and divergence from ancestral genes (31). This overprinting hypothesis has gained recent attention with respect to the origins of genetic novelty (32). Aside from overprinting, two related and previously unknown mechanisms for protein novelty from alternative reading frames—gene remodeling

* Corresponding author e-mail: zachary.ardern@sanger.ac.uk

and pairs of compensatory frameshift mutations—have recently been elucidated (33, 34); these are discussed later in this study.

An overlapping gene pair consists of a reference frame sequence and a sequence encoded in one of the five overlapping alternative reading frames (Figure 1A). Elsewhere, these have been referred to as the (pre-existing) "mother gene" and (younger, overprinted) "daughter gene" by analogy to mother and daughter cells in reproduction (8). The phenomenon studied here, however, is broader, in that it also includes alternative frame sequences that are not functional genes. Thus, we refer simply to reference and alternative reading frames. Some properties of the sequences in alternative reading frames are dictated by the structure of the standard genetic code. For a novel protein sequence encoded in an alternative frame of an existing protein-coding gene, some properties are determined by its position in relation to the existing reference frame sequence. The details will depend on the genetic code table mapping codons to amino acids. For instance, with the standard genetic code the two alternative reading frames on the same strand as a given "reference frame" sequence tend to encode amino acids which preserve the hydrophobicity profile of the reference frame's protein sequence (35). This finding has been argued to be an artefact of selection on the code for robustness to point mutations (36), so we have not listed it here as an "optimality" of the code, but it may deserve further attention. We have previously shown that both purported optimalities and a lack of optimality are sensitive to code sets, threshold choices, and the combination of properties used (1).

The widespread assumption that frameshifts have no biological use is also called into question by some of the studies discussed here and by our results. Aside from the tendency for preservation of hydrophobicity in same strand alternative reading frames, the structure of the standard genetic code also creates a strong tendency for amino acids of opposite hydrophobicities to be encoded by the codons directly in antisense to each other (37). This property may perhaps assist in creating a template of structural elements in antisense open reading frames (38). Other aspects of evolution in this directly antisense frame "a3" and corresponding protein properties have also been explored (39, 40). Importantly, not all properties of overlapping genes are simply a result of code structure, as genuinely translated overlapping genes are just a small and presumably

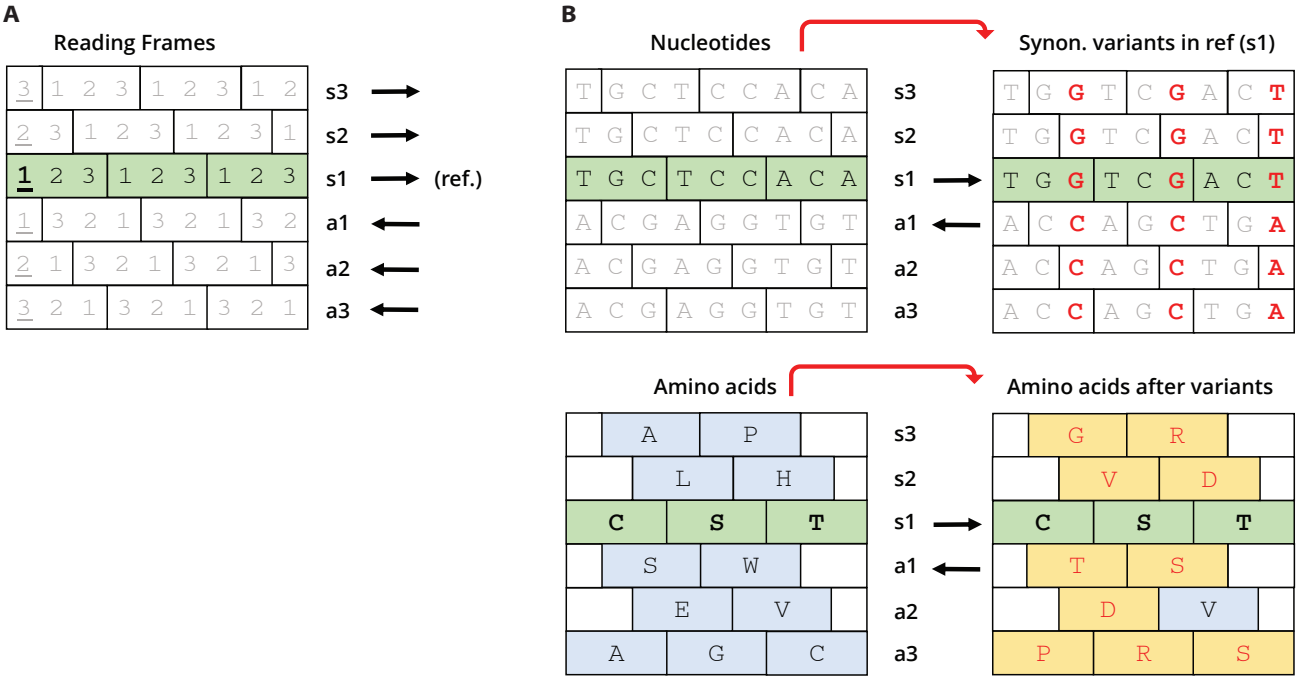

Figure 1. (A) The five alternative reading frames, named according to a shortened form of the schema used in Wei and Zhang (42). In this schema s = sense, a = antisense, and the number represents the codon position of the alternative frame that corresponds to position 1 in the reference frame; the relevant codon position numbers are underlined here. (B) Example of "collateral effects" in alternative frames following synonymous mutations in the reference frame. Variants at the nucleotide level and changes in amino acids encoded by alternative reading frames are shown in red and light yellow; the original amino acids encoded in the alternative frames are shown in blue.

non-random subset of the sequences found in alternative reading frames. For instance, it has been shown that the higher intrinsic disorder of overlapping versus nonoverlapping proteins is not due simply to code structure (41).

The arrangement of the standard genetic code could very plausibly be different, resulting in different properties in alternative reading frames. There are variant genetic codes across diverse taxonomic groups, although these minor variants extant today appear all to be derivative of the standard code (43, 44). A triplet code encoding the 20 canonical amino acids or a subset of these can be arranged in more than 10^{80} different ways. The actual arrangement has been shown to be somewhat "optimal" in comparison with alternative possible codes, across various biologically useful features, including robustness to point mutations (45), termination after a frameshift, and the incorporation of additional noncoding information within protein-coding sequences (46). We summarized some of this literature and analyzed a few properties in a previous study (1). The choice of the 20 amino acids is also near optimal in its coverage of physico-chemical space (47-49).

Intriguingly, the structure of the code is not just beneficial for biological functions such as minimizing the effect of mistranslation, but it also facilitates evolution. Minimizing mutation effect size promotes adaptation, as seen, e.g., in Fisher's Geometric Model of adaptation (50, 51). Beyond this, study of an empirical fitness landscape showed that the code helps ensure that mutations are both depleted in deleterious variants and enriched for adaptive variants (52). This optimizes the exploration of functional variants at intermediate time scales, as shown by a later study using a larger experimentally derived fitness landscape (53). While investigating the potential multidimensional optimality of the code suggested by some of the studies cited here, we discovered that synonymous mutations in a reference reading frame have remarkably similar average collateral mutation effects across at least four out of five alternative reading frames (1). In this study we redo this calculation for all five alternative reading frames and investigate potential ramifications of this finding by means of a very simple evolutionary model.

Evolution by natural selection can be visualized as a process of climbing peaks in a fitness landscape. The topology of real fitness landscapes is still being investigated (54–57), but it has been shown they are often "rugged," meaning that there are distinct local fitness peaks in addition to the global peak. Such ruggedness can limit adaptive evolution, depending on the height of the local peaks and their proximity to a global peak. Considerations of fitness landscapes in the context of individual genes generally concern situations where a functional sequence (e.g., a protein-coding gene) is undergoing adaptation. Here we are interested in processes of sequence exploration more generally, particularly the origin of new functional sequences. This requires not just minor adaptation, but also a wider exploration of sequence space—thus, we consider two different types of evolutionary events. The first are small-effect "conservative" mutations, which are more likely to shift a sequence toward a fitness peak, in line with standard Darwinian processes as seen for instance in Fisher's Geometric Model (58, 59). The second are large effect explorative mutations, which can assist with moving between fitness peaks (local maxima). These are required to sample disparate regions of sequence space and to prevent populations from getting stuck in small local maxima.

Methods

The methods for the calculation of alternative frame average mutation effect size in a given genetic code are reported in our previous study (1). According to the standard approach in the genetic code optimality literature (60, 61), "mutation effect size" is measured as the square of the difference in polar requirement between two amino acids. The value of interest here is the average effect size in alternative reading frames of mutations that are synonymous in a reference reading frame, i.e., the average effect size (square of the difference in polar requirement) for the amino acids encoded in alternative frame codons following synonymous mutations in reference reading frame codons (Figure 1B). In the previous study the directly antisense frame "a3" (Figure 1A) had effect sizes that were approximately 20 times larger, since it overlaps a single codon in the reference frame while other alternative frames overlap dicodons. This discrepancy between frames was due to the incorporation of weights for both (i) how frequently each amino acid is used in a genome and (ii) the number of conservative mutations that are possible for each amino acid in the reference frame. Removing one of the two weights makes the "a3" frame comparable with the other reading frames. Here the weighting for the number of possible conservative mutations is removed. The average mutation effect size in the alternative reading frames was calculated in a large set of 10^7 possible genetic code tables sharing the block structure of the standard genetic code, including the standard genetic code itself.

The model in this study investigates whether an average mutation effect size (i.e., an average step size in our model) can be chosen so as to maximize the number of sequences ending up in a large fitness peak. We represent sequence space as a 2D surface with periodic boundary conditions (i.e., the space in the model loops back on itself so there is no "edge" to the map). Particles represent sequences, which shift in sequence space over the course of generations, due to the effects of mutations (Figure 2A). Sequences are altered (particles move

around the map) in one of two ways: in steps that are either “conservative” or "explorative" (Figure 2B; Table 1). The “conservative” mutations have a small step size s_c and have an inbuilt higher chance of moving the sequence to higher fitness. That is, if the sequence is within a fitness peak region, these mutations are biased toward moving toward the center of the peak, modeling the influence of the filter of natural selection (positive selection) on such mutations. The direction of the larger explorative mutations on the other hand is not influenced by fitness, and thus for these mutational vectors all angles are equiprobable regardless of the particle’s location in the map. These represent mutations with a large effect that happen to evade purging by natural selection.

The direction of movement of the steps of either step size, s_c or s_e, is stochastic. For conservative mutations the probability of moving in each possible direction depends on the relative fitness in this direction compared with that of others. In detail, the possible directions for mutational steps are discretized into N equiangular directions, with $N = 100$ used throughout this study. The fitness f_i value after each possible step move is calculated, and from this the minimum fitness after a movement, f_{min}, is determined. The probability of a shift in any given direction, p_i, is calculated using this as in Equation 1. The addition of +1 to the numerator ensures that $p_i > 0$, and the denominator is a normalization so that $\sum p_i = 1$. Thus, the equation entails that the probability of a move in any direction is effectively determined by the difference between the fitness value after moving in that direction with step size s_c (i.e., f_i) and the minimum fitness value possible across all mutations with step size of s_c (i.e., f_{min}). A probability value $p_i > 0$ for each position is required in order to properly simulate the boundary areas of a fitness peak. Without the addition of a constant to the numerator, points with the lowest fitness value cannot be moved to, so a particle close to a fitness peak would get absorbed by it with a rate independent of its fitness value, as the entire area outside the peak has the lowest fitness value of zero.

$$p_i = \frac{1 + f_i - f_{min}}{\sum_{j=0}^{N} (1 + f_j - f_{min})}$$

Equation 1:
Probability, for conservative mutations, of the mutational step being in any given direction. Explorative mutations have no directional bias.

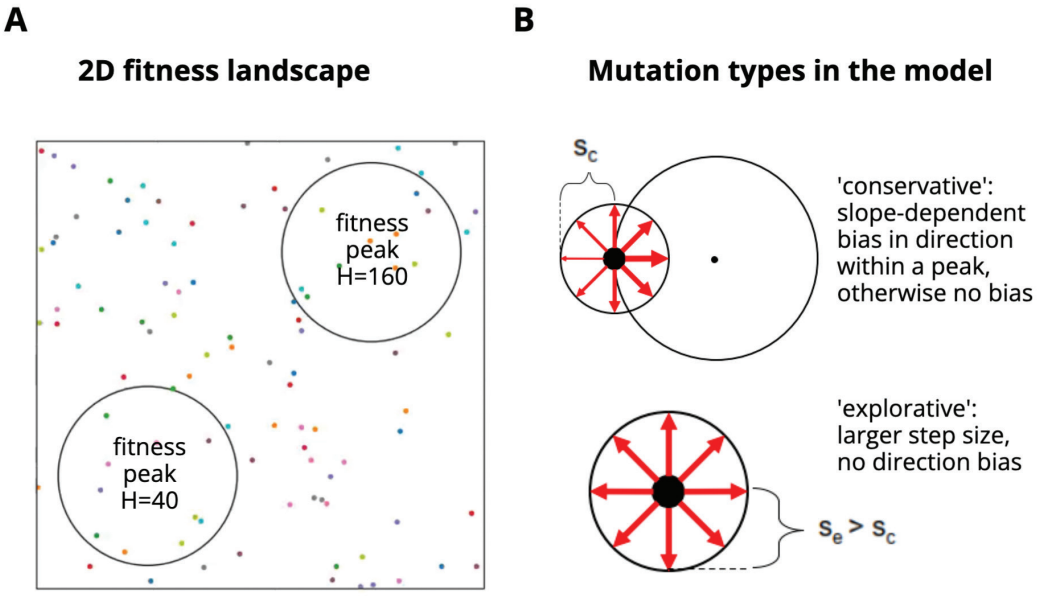

Figure 2. (A) In this 2D representation of the model, sequence space is represented with circles corresponding to fitness peaks, and an initial distribution of sequences is represented as colored points. Each fitness peak is a symmetric cone with radius 0.2 in a normalized sequence space of size 1 x 1. The top right peak’s fitness value (height, H) is 160 and the lower left peak has H = 40. (B) Illustration of conservative and explorative mutations in the model—the most likely direction of conservative mutations is affected by whether the sequence (the point particle) is situated within a peak, while explorative mutations are not, and the magnitude of the conservative mutation step size is smaller than the magnitude of explorative mutation step size (i.e., $s_e > s_c$).

Table 1. Summarized properties of the two types of evolutionary "step" within the model.

Mutation type	Step size	Biased toward higher fitness?	Probability of this mutation
conservative	s_c	Yes, if within a peak	p_c
explorative	s_e	No	1-p_c

Results

First, following up on the previously reported results (1) whereby the standard genetic code appears to minimize differences in the average alternative-frame effect size for synonymous mutations, we confirmed that this property is very rare among alternative genetic codes. Among 10^7 codes sharing the block structure of the standard genetic code, only 0.77% of codes had an equivalent or greater similarity in average collateral effect size across frames compared to the standard genetic code (Figure 3, Supplementary Figure 1). The measure used to determine similarity is the standard deviation σ_D between mutation effect values D_c in different reading frames. We find that the standard deviation of the mutation effect values between the different reading frames in the standard genetic code is very low. That is, compared with other possible codes, the standard code ensures that average mutation effect sizes are remarkably similar across the alternative frames.

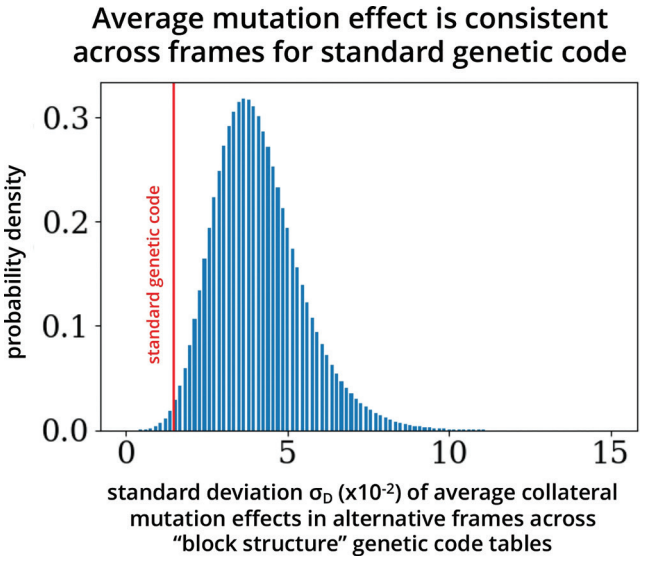

Figure 3. The standard deviation σ_D of mutation effect values (D_c) across all alternative frames calculated for 10^7 alternative codes. The position of the standard genetic code among the alternative code set is shown with the vertical red line (only 0.77% of other "block" codes have equivalent consistency across frames). This style of figure is standard in the code optimality literature.

For the main part of this study, we consider a simplification of the effects of large and small mutations in a rugged fitness landscape. Both conservative mutation and larger scale evolutionary exploration of sequence space are potentially advantageous for the evolution of novel genes. Larger mutations allow exploring sequence space in order to find functional regions or improve fitness. Conservation allows functional sequences to be maintained in a population without being immediately degraded (i.e., reduced in fitness) by new mutations, and conservative mutations also facilitate small-scale adaptation within a fitness peak, as discussed above regarding Fisher's Geometric Model.

We hypothesize that for a given rugged fitness landscape, there is an optimal trade-off between evolvability (or exploration) and robustness in mutation effects that will maximize the fitness of a population of novel sequences, i.e., facilitate finding high fitness peaks. Whether the standard genetic code has actually achieved an optimal value is not addressed here—this study is groundwork for further investigation of this issue. In support of the hypothesis, we present results from a simple model with two kinds of evolutionary processes, namely "explorative" and "conservative," with sequences evolving on a fitness landscape with two broad fitness peaks of different heights, as described in more detail above in Methods.

The model illustrates the expected dynamics, where sequences accumulate in the larger fitness peak over time, leaving the smaller peak (Figure 4A). Stochastic fluctuations in the proportion of sequences in the smaller fitness peak are larger than for the higher peak, reflecting the smaller peak's lower capacity to retain sequences. If the probability of mutations being conservative, p_c, is decreased, then stochastic fluctuations in the distribution of sequences across peaks are larger, fewer sequences are found in the high peak at most time points, and the average fitness value of sequences is lower, as many sequences are not in the high peak (Figure 4B). Population fitness over the medium to long term can thus be optimized by fine-tuning the degree of stochasticity (due to large-effect mutations) to a point at which as many sequences as possible reach the high peak and are retained in it.

The model described so far uses two different step sizes and calculates a ratio of how often each occurs. In order to use this model to test whether an average step size can be chosen so as to optimize sequence average fitness, the model is run to obtain the population's average fitness values over a range of the parameters for step size and probability of each kind of mutation, as shown in Figure 5. Within the parameter space,

we observe three qualitatively different regions, labeled as I, II, and III. In region I, conservative steps predominate, and sequences remain in whichever fitness peak they are in, whether the high or low peak. In region III, the opposite tendency is observed and neither of the two peaks can retain sequences over the long term, i.e., stochastic fluctuations dominate. In the small high fitness area between I and III, i.e., region II, sequences are conserved in the higher but not the lower peak, as was observed in Figure 4A. The average mutation step size s is calculated with the equation $s = p_c \times s_c + (1 - p_c)s_e$. Fitting this equation (black line) to region II shows that s can be chosen so as to give a close approximation to the function which optimizes fitness. In other words, we observe a non-trivial similarity between the fitness-optimizing set of parameters and the result of a constant average mutation step size.

It may not be immediately clear how this model relates to the biological reality of alternative frames. Different reading frames are expected to have very different distributions of effect sizes for conservative mutations. For example, the "a2" (Figure 1A) frame is mostly a very conservative reading frame (62) —we would not expect it to have a similar average collateral mutation effect size to the other frames. In order to have a similar average mutation effect size as other frames, among the possible collateral mutation effects there must be rare variants with large effect sizes. Its mutation effect size distribution is expected to vary substantially from the other reading frames, with the other frames having a more even distribution of effect sizes. After calculating the effects for each alternative frame, we indeed observe that the a2 frame has more possible variants of large effect, balancing out the generally "conservative" nature of variant effects (Supplementary Figure 2). Thus, every reading frame occupies a different point on the black line (where there is a constant average mutation effect) illustrated in Figure 5. In summary, if the right average mutation size is instantiated each frame could optimize the average fitness in its own way. The optimal region II is also very thin, so the average mutation step size must be very similar across reading frames as observed in the SGC. The details of this and its impacts on the biological and evolutionary use of each alternative frame deserve further attention.

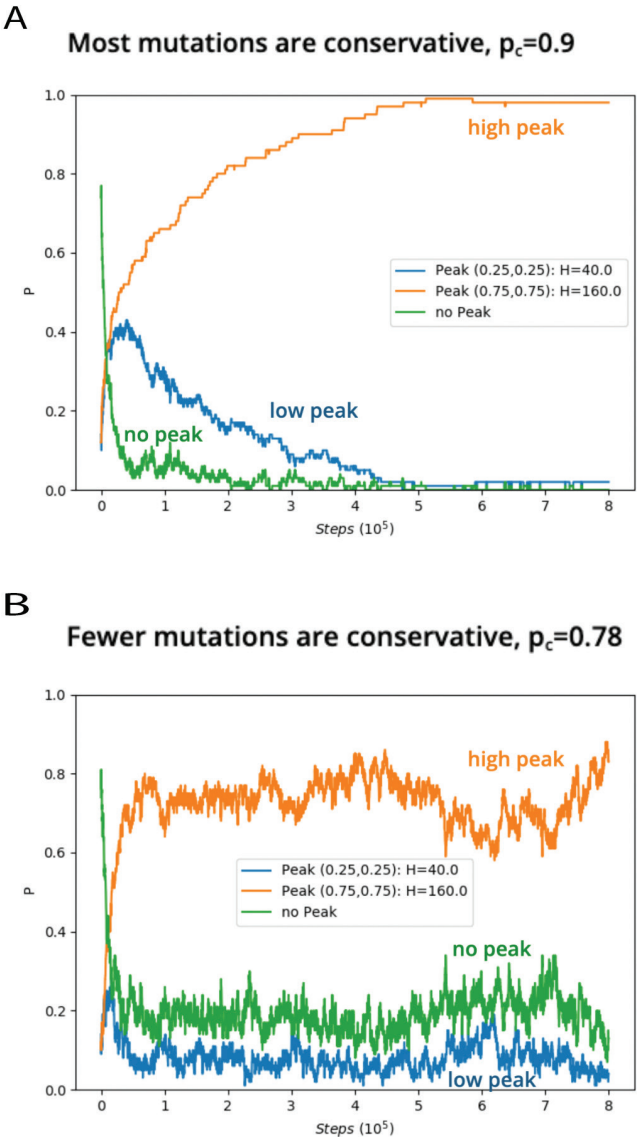

Figure 4. The proportion P of all sequences in one of the two peaks or outside any peak across 8×10^5 mutation steps. The data was created from 100 sequences, s_c = 0.001 and s_e = 0.01. A: With an approximately optimal value of p_c = 0.9 most sequences eventually end up in the higher peak (H=160) with some stochastic fluctuations. B: When fewer mutations are conservative (p_c = 0.78), sequences accumulate in the higher peak faster, but with larger stochastic fluctuations, and the population as a whole does not reach high fitness. The effect of other parameter values is shown in Figure 5.

Discussion

We present a simple evolutionary model examining the previously discovered (1) property of a remarkably consistent collateral effect size of synonymous mutations across alternative reading frames. We have also updated the calculation of mutation effect sizes across alternative frames for sets of alternative codes. Our simple model shows that for a given simple fitness landscape the average mutation step size is able to be fine-tuned so as to optimize population fitness. This is only a toy model, presented as a proof of concept, laying the groundwork for further investigation. Multiple aspects of the model can be called into question for

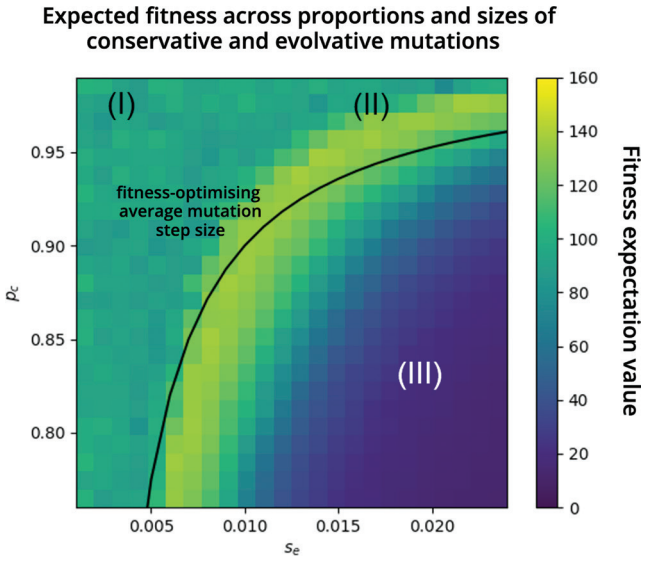

Figure 5. The average fitness value of 100 sequences after 8×10^5 mutations for different values of conservative mutation probability p_c and explorative mutation step size s_e. The conservative mutation step size is fixed to $s_c = 0.001$. Fitness expectation values fall into three regions. In (I), sequences are retained in both peaks, in (II), sequences are restricted to the higher peak, and in (III), sequences are not retained in either peak. If we keep the average mutation effect size at a constant value (black line) we find almost the same functional relation between p_c and s_e as that describing region (II) which optimizes the expected fitness.

biological realism. However, the key finding that the function for average step size so closely approximates the function for optimizing average sequence fitness appears to give real insight into evolutionary dynamics. After discussing ways in which this work can be developed or explored further, we will unpack some implications of the work and comment on the broader field of study of alternative reading frames.

First, we note that the initial observation of similar collateral effects across alternative frames adds another potential optimality to the growing list of interesting properties of the standard genetic code. Whether this is an artifact of some other property of the code or the method used for calculation and how to interpret the finding if it does hold up both require further attention. It is important to note that some other putatively optimal properties have been shown not to be independently optimal and thus are perhaps artefacts of already-known facets of code structure. In addition to the putative frameshift optimality addressed above (36), a prominent study in *Science* claimed that the code is near optimal for resource conservation (63), but this claim has been subjected to rigorous critique in two response papers (64, 65). The idea that the structure of the code may be a result of selection in general remains controversial (1, 66–68), but optimality is a distinct question from that of historical process.

Future research on this topic could include methodological improvements, more realistic modeling, investigation of OLG evolution in real biological data, and further investigation of the origins of the remarkable structure of the standard genetic code. In terms of improving biological realism in the model, a number of steps can be taken. If the observed results hold across both more realistic models of evolving populations and diverse fitness landscapes this will support our claim; in principle, we believe that they should. Other methods of calculating effect sizes across frames could be investigated. Our results should also be integrated with previous research on fitness landscapes—it seems likely that the phenomenon we report where an average mutation size can optimize fitness has been reported in a different context, but if so, we have not found it. Finally, regarding further developments, investigating sequence data from evolving OLGs is a potential new research field of its own. Until recently, a large-scale analysis was not possible, but improved methods for detecting overlapping genes' protein expression such as ribosome profiling (69, 70) or detecting sequence features associated with OLGs (71–74) will allow studies across taxa.

The model can be seen as illustrating the process of finding function amidst the vast hyperastronomical (75) ocean of possible protein sequence space, the large majority of which is not functional even by relatively optimistic estimates (76). Interpreted in this way, the model suggests that the right mutation effect size will help to maximize a population's ability to find and retain new functions. Alternatively, instead of functionless regions of sequence space, the regions outside the peaks in the model can be conceived of as illustrating neutral networks across sequence space (i.e., networks of similar sequences with the same function), with perhaps a low level of functionality. In either case, a particle moving into a peak represents finding functional novelty in sequence space.

Whether evolvability or robustness is more important in real life depends on various parameters of the evolutionary processes in which alternative frame coding is involved. The idea of a trade-off between robustness and evolvability in the genetic code's structure, facilitating the search for functional sequences, has been proposed before in the context of normal adaptive evolution (53). Our results can be seen as an extension of this. It is also possible however that the apparent trade-off between robustness and evolvability in mutation effect size is better conceived of in another way than in terms of directly optimizing fitness. For instance, as new

sequences are sampled by evolution, small mutations allow a new sequence to be sampled relatively unchanged, while large mutations cause a jump to new regions of sequence space, and both may be needed to optimally sample the total space and find new functions. Thus, a more relevant trade-off may involve optimizing the time taken for sequence search rather than long-term population fitness, but we have not explored this here.

How does the potential origin of protein sequence novelty from alternative reading frames concretely impact biological reality? Recent findings on the role of alternative reading frames in protein origins is beginning to shed light on this. Alternative reading frames can potentially be used for protein novelty in at least three different ways: overprinting, remodeling, and frameshift mutations. The first to be discovered and investigated was the process of "overprinting," where an out-of-frame open reading frame becomes translated and is retained due to some functional advantage. Conceivably, the resulting overlapping gene pair could then be copied, and the homologue of the original gene then pseudogenized (30). After some time, this would leave essentially no trace of the original gene, and may explain the origin of some "orphan genes" — but to our knowledge no specific examples of this have yet been demonstrated. In eukaryotes, a related mechanism termed "mosaic translation" has been hypothesized but to our knowledge not demonstrated (77) —this is where different parts of a spliced RNA transcript are read in different reading frames, producing "mosaic" proteins. The second process termed remodeling has recently been demonstrated to play a nontrivial role in the origin of genes in *E. coli*. This is where gene fusion occurs between either part of a gene-read in-frame and a frameshifted sequence or multiple frameshifted sequences (33). A third process has also very recently been demonstrated, namely, "mutually compensating frameshift mutations," where two sequential frameshift mutations result in a partial frameshift that is tolerated in a protein sequence and opens up new sequence region for evolutionary exploration (34). More generally, if partially frameshifted sequences are tolerated, they constitute new protein sequences. All of these mechanisms bear some similarity to another process highlighted recently, where stochastic stop-codon read-through allows some noncoding sequences after stop codons to contribute to protein novelty (78).

Whether overprinting, remodeling, or frameshifting predominates is unclear, and the distribution is likely to be taxon dependent. In general, these processes have received little attention, likely, at least, partly due to a perceived limitation from evolutionary constraint (25, 62). Evolutionary constraint in OLGs has not yet been studied in much detail for specific gene pairs, apart from the case of overlapping genes in HIV-1. In this virus it has been shown first that constrained regions of overlapping genes in a pair are organized so as not to overlap (79) and second that even when domains do overlap, functionally constrained residues in one protein are encoded overlapping more mutable codons in the other protein (80). Further, constraint can be conceived either negatively or positively. On the negative side, a study of constraints across reading frames concluded that the most constrained frame (a2 in Figure 1A, often "-2" elsewhere) is likely to be very rare in nature given the constraints in codons permitted in that frame (62). On the other hand, constraint could bias overlapping frames toward functional sequences, acting as an approximate template (i.e., ensuring that some semblance of protein structure is encoded in many of these sequences).

As described above, evidence for this actually being the case includes potential structure-promoting biases in both same-strand (35) and antisense (38) frames. It has already been shown that constraints from the structure of the genetic code facilitate Darwinian evolution (52), so it is reasonable to look for more cases of this.

In presenting the provocative hypothesis that this aspect of the structure of the standard genetic code may be "adaptive," we are not making any specific claim about the processes behind the origin of the code's structure. We have discussed some aspects of this in a previous publication (1). In this context "adaptive" (present usefulness) should not be conflated with "having been adapted" (historical process of fitting for a use). Similarly, the fact of being "optimal for X" should not be conflated with "having been optimized for X," and does not need to imply that the feature originated through a process of natural selection. Many aspects of biology are functional, may appear very "apt" for their functions, and can even be essential, without being a direct result primarily of selection. These include RNA secondary structures (81) and various examples of biochemical complexity (82). A related concept, although one that has accumulated a lot of conceptual baggage, is a "spandrel" (83) The idea that the code has really been selected in order to be optimal across multiple parameters is perhaps implausible, given both the inherent difficulties in evolving any functional code (any change to the code will change multiple messages) and the limited timespan available for code evolution between the origin of life and the last universal common ancestor. The optimality of the code could be an example of what has been termed evolutionary inherency (84), where structures developed early in evolution end up being put to different functional use much later (such as seen in various components of animal nervous systems). Regardless of historical causes, the structure of the code and the uses it has in modern organisms are worth investigating.

In summary, our key results are found in Figure 3 and Figure 5. First, the standard genetic code has the property of a remarkably consistent average mutation effect size (measured in terms of amino acid polarity) in alternative reading frames following synonymous mutations in the reference frame. Second, we investigate a way in which this apparent optimality could be biologically useful with a simple evolutionary model. In our model it is possible to optimize fitness across different step sizes of large "explorative" mutations, by choosing the proportion of mutations that are explorative (versus conservative) so that the overall average mutation effect size is maintained at a particular optimal value. Analogously, the constant average mutation effect size across reading frames, despite other across-frame differences in the distributions of mutation effect sizes, may assist in finding functional sequences in the alternative frames.

The main point we hope that the reader takes away from this study is the exciting potential for further research regarding both the contribution of the genetic code to evolvability and the origin of protein novelty from alternative reading frames.

Acknowledgments

This manuscript is derived from work conducted as part of the PhD thesis of SW, submitted for examination at the Technical University of Munich and supervised by ZA. The model was presented in a poster at the conference “The Physics of Evolution” at the Francis Crick Institute, London, in July 2019. Thanks to everyone who gave feedback on the project and to Siegfried Scherer and Klaus Neuhaus for supervision of the overall PhD and related projects.

Funding

Funding of work conducted in the research group of Prof. Siegfried Scherer at the Technical University of Munich was obtained from the Bavarian State Government and the National Philanthropic Trust. Work conducted at the Wellcome Sanger Institute was supported by the Wellcome Trust, grant [108413/A/15/D]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to the Author Accepted Manuscript, available at BioRxiv https://doi.org/10.1101/2022.03.22.485379.

References

1. Wichmann S, Ardern Z. Optimality in the standard genetic code is robust with respect to comparison code sets. Bio Systems. 2019 November;2019;185:104023. doi:10.1016/j.biosystems.2019.104023
2. Barrell BG, Air GM, Hutchison CA 3rd. Overlapping genes in bacteriophage phiX174. Nature. 1976;264(5581):34-41. doi:10.1038/264034a0
3. Firth AE, Brierley I. Non-canonical translation in RNA viruses. J Gen Virol. 2012;93(Pt 7):1385-1409. doi:10.1099/vir.0.042499-0
4. Cassan E, Arigon-Chifolleau AM, Mesnard JM, Gross A, Gascuel O. Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. Proc Natl Acad Sci USA. 2016;113(41):11537-11542. doi:10.1073/pnas.1605739113
5. Affram Y, Zapata JC, Gholizadeh Z, Tolbert WD, Zhou W, Iglesias-Ussel MD, Pazgier M, Ray K, Latinovic OS, Romerio F. The HIV-1 antisense protein ASP is a transmembrane protein of the cell surface and an integral protein of the viral envelope. J Virol. 2019;93(21):e00574-19. doi:10.1128/JVI.00574-19
6. Nelson CW, Ardern Z, Goldberg TL, Meng C, Kuo CH, Ludwig C, Kolokotronis SO, Wei X. Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic. Elife. 2020;9:e59633. doi:10.7554/eLife.59633
7. Firth AE. A putative new SARS-CoV protein, 3c, encoded in an ORF overlapping ORF3a. J Gen Virol. 2020;101(10):1085-1089. doi:10.1099/jgv.0.001469
8. Kreitmeier M, Ardern Z, Abele M, Ludwig C, Scherer S, Neuhaus K. Spotlight on alternative frame coding: Two long overlapping genes in Pseudomonas aeruginosa are translated and under purifying selection. iScience. 2022;25(2):103844. doi:10.1016/j.isci.2022.103844
9. Zehentner B, Ardern Z, Kreitmeier M, Scherer S, Neuhaus K. Evidence for numerous embedded antisense overlapping genes in diverse E. coli strains. bioRxiv. 2020. Available from: https://doi.org/10.1101/2020.11.18.388249
10. Ardern Z, Neuhaus K, Scherer S. Are Antisense Proteins in Prokaryotes Functional?. Front Mol Biosci. 2020;7:187. doi:10.3389/fmolb.2020.00187
11. Meydan S, Vázquez-Laslop N, Mankin Alexander S. Genes within genes in bacterial genomes. Microbiology Spectrum. 2018;6(4). Available from: https://doi.org/10.1128/microbiolspec.RWR-0020-2018
12. Hücker SM, Vanderhaeghen S, Abellan-Schneyder I, Scherer S, Neuhaus K. The novel anaerobiosis-responsive overlapping gene ano is overlapping antisense to the annotated gene ECs2385 of Escherichia coli O157:H7 Sakai. Front Microbiol. 2018;9:931. doi:10.3389/fmicb.2018.00931
13. Vanderhaeghen S, Zehentner B, Scherer S, Neuhaus K, Ardern Z. The novel EHEC gene asa overlaps the TEGT transporter gene in antisense and is regulated by NaCl and growth phase. Sci Rep. 2018;8(1):17875. doi:10.1038/s41598-018-35756-y
14. Gelsinger DR, Dallon E, Reddy R, Mohammad F, Buskirk AR, DiRuggiero J. Ribosome profiling in archaea reveals leaderless translation, novel translational initiation sites, and ribo-

some pausing at single codon resolution. Nucleic Acids Res. 2020;48(10):5201-5216. doi:10.1093/nar/gkaa304
15. Loughran G, Zhdanov AV, Mikhaylova MS, Andreev DE. Unusually efficient CUG initiation of an overlapping reading frame in POLG mRNA yields novel protein POLGARF. 2020;117(40):24936-24946. Available from: https://doi.org/10.1073/pnas.2001433117
16. Khan YA, Jungreis I, Wright JC, Mudge JM, Choudhary JS, Firth AE, Kellis M. Evidence for a novel overlapping coding sequence in POLG initiated at a CUG start codon. BMC Genet. 2020;21(1):25. doi:10.1186/s12863-020-0828-7
17. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Gonzalez JM, Magrane M, Martinez T, Schulz JF, Yang YT, Alba MM, et al. A community-driven roadmap to advance research on translated open reading frames detected by Ribo-Seq. bioRxiv. 2021. Available from: https://doi.org/10.1101/2021.06.10.447896
18. Cao X, Khitun A, Luo Y, Na Z, Phoodokmai T, Sappakhaw K, Olatunji E, Uttamapinant C, Slavoff SA. Alt-RPL36 downregulates the PI3K-AKT-mTOR signaling pathway by interacting with TMEM24. Nat Commun. 2021;12(1):508. doi:10.1038/s41467-020-20841-6
19. Wright BW, Yi Z, Weissman JS, Chen J. The dark proteome: translation from noncanonical open reading frames. Trends Cell Biol. 2022;32(3):243-258. doi:10.1016/j.tcb.2021.10.010
20. Szekely M. Triple overlapping genes. Nature. 1978;272(5653):492.
21. Siegel AF, Fitch WM. Degeneracy when DNA codes for overlapping genes. Mathematical Biosciences. 1980;49(1):1-16. Available from: https://doi.org/10.1016/0025-5564(80)90107-8
22. Smith TF, Waterman MS. Overlapping genes and information theory. J Theoret Biol. 1981;91(2):379-380.
23. Yockey HP. Rebuttal of 'overlapping genes and information theory.' J Theoret Biol. 1981;91(2):381-382.
24. Miyata T, Yasunaga T. Evolution of overlapping genes. Nature. 1978;272(5653):532-535.
25. Yockey HP. Do overlapping genes violate molecular biology and the theory of evolution? J Theoret Biol. 1979;80(1):21-26.
26. Kolata GB. Overlapping genes: more than anomalies? Science. 1977;196(4295):1187-1188.
27. Wright BW, Molloy MP, Jaschke PR. Overlapping genes in natural and engineered genomes. Nat Rev Genet. 2022;23(3):154-168. doi:10.1038/s41576-021-00417-w
28. Brandes N, Linial M. Gene overlapping and size constraints in the viral world. Biol Direct. 2016 May;11:26.
29. Vakirlis N, Carvunis AR, McLysaght A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. Elife. 2020;9:e53500. doi:10.7554/eLife.53500
30. Keese PK, Gibbs A. Origins of genes: "big bang" or continuous creation?. Proc Natl Acad Sci USA. 1992;89(20):9489-9493. doi:10.1073/pnas.89.20.9489
31. Ohno S. Evolution by gene duplication. Berlin: Springer; 1970.
32. Carter CW. Simultaneous codon usage, the origin of the proteome, and the emergence of de-novo proteins. Cur Opin Struct Biol. 2021;68:142-148.
33. Watson AK, Lopez P, Bapteste E. Hundreds of out-of-frame remodeled gene families in the escherichia coli pangenome. Mol Biol Evol. 2022;39(1):msab329. Available from: https://doi.org/10.1093/molbev/msab329
34. Biba D, Klink G, Bazykin GA. Pairs of mutually compensatory frameshifting mutations contribute to protein evolution. Mol Biol Evol. 2022;39(3):msac031. Available from: https://doi.org/10.1093/molbev/msac031
35. Bartonek L, Braun D, Zagrovic B. Frameshifting preserves key physicochemical properties of proteins. Proc Natl Acad Sci USA. 2020;117(11):5907-5912.
36. Xu H, Zhang J. On the origin of frameshift-robustness of the standard genetic code. Mol Biol Evol. 2021a;38(10):4301-4309. doi:10.1093/molbev/msab1642021a
37. Blalock JE, Smith EM. Hydropathic anti-complementarity of amino acids based on the genetic code. Biochem Biophys Res Comm. 1984;121(1):203-207.
38. Zull JE, Smith SK. Is genetic code redundancy related to retention of structural information in both DNA strands? Trends Biochem Sci. 1990;15(7):257-261.
39. Konecny J, Eckert M, Schöniger M, Hofacker GL. Neutral adaptation of the genetic code to double-strand coding. J Mol Evol. 1993;36(5):407-416.
40. Blalock JE. Complementarity of peptides specified by 'sense' and 'antisense' strands of DNA. Trends Biotechnol. 1990;8(6):140-144.
41. Willis S, Masel J. Gene birth contributes to structural disorder encoded by overlapping genes. genetics. 2018;210(1):303-313. doi:10.1534/genetics.118.301249
42. Wei X, Zhang J. A simple method for estimating the strength of natural selection on overlapping genes. Genome Biol Evol. 2015;7(10):381-390. Available from: https://doi.org/10.1093/gbe/evu294
43. Osawa S. Evolution of the genetic code. Oxford: Oxford University Press; 1995.
44. Freeland SJ, Knight RD, Landweber LF, Hurst LD. Early fixation of an optimal genetic code. Mol Biol Evol. 2000;17(4):511-518. doi:10.1093/oxfordjournals.molbev.a026331
45. Freeland SJ, Hurst LD. The genetic code is one in a million. J Mol Evol. 1998;47(3):238-248.
46. Itzkovitz S, Alon U. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res. 2007;17(4):405-412.
47. Ilardo M, Meringer M, Freeland S, Rasulev B, Cleaves HJ 2nd. Extraordinarily adaptive properties of the genetically encoded amino acids. Sci Rep. 2015;5:9414. doi:10.1038/srep09414
48. Ilardo M, Bose R, Meringer M, Rasulev B, Grefenstette N, Stephenson J, Freeland S, Gillams RJ, Butch CJ, Cleaves HJ 3rd. Adaptive properties of the genetically encoded amino acid alphabet are inherited from its subsets. Sci Reports. 2019;9(12468). Available from: https://doi.org/10.1038/s41598-019-47574-x
49. Mayer-Bacon C, Freeland SJ. A broader context for understanding amino acid alphabet optimality. J Theo Biol. 2021 July;520:110661.

50. Freeland SJ. The Darwinian genetic code: An adaptation for adapting? Genet Program Evolvable Mach. 2002;3(2):113-127. Available from: https://doi.org/10.1023/A:1015527808424
51. Zhu W, Freeland SJ. The standard genetic code enhances adaptive evolution of proteins. J Theoret Biol. 2006;239(1):63-70.
52. Firnberg E, Ostermeier M. The genetic code constrains yet facilitates Darwinian evolution. Nucleic Acids Res. 2013;41(15): 7420-7428.
53. Tripathi S, Deem MW. The standard genetic code facilitates exploration of the space of functional nucleotide sequences. J Mol Evol. 2018;86(6):325-339.
54. Richter H, Engelbrecht A, editors. Recent advances in the theory and application of fitness landscapes. Berlin: Springer; 2014.
55. de Visser JA, Krug J. Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet. 2014;15(7):480-490. doi:10.1038/nrg3744
56. Payne JL, Wagner A. The causes of evolvability and their evolution. Nat Rev Genet. 2019;20(1):24-38.
57. Chen JZ, Fowler DM, Tokuriki N. Environmental selection and epistasis in an empirical phenotype-environment-fitness landscape. Nat Ecol Evol. 2022;6(4):427-438. doi:10.1038/s41559-022-01675-5
58. Tenaillon O. The utility of Fisher's geometric model in evolutionary genetics. Annu Rev Ecol Evol Syst. 2014;45:179-201. doi:10.1146/annurev-ecolsys-120213-091846
59. Fisher RA. The genetical theory of natural selection. Oxford: Clarendon Press; 1930. Available from: https://doi.org/10.5962/bhl.title.27468
60. Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC. On the fundamental nature and evolution of the genetic code. Cold Spring Harb Symp Quant Biol. 1966;31:723-736. doi:10.1101/sqb.1966.031.01.093
61. Buhrman H, van der Gulik PT, Kelk SM, Koolen WM, Stougie L. Some mathematical refinements concerning error minimization in the genetic code. IEEE/ACM Trans Comput Biol Bioinform. 2011;8(5):1358-1372. doi:10.1109/TCBB.2011.40
62. Lèbre S, Gascuel O. The combinatorics of overlapping genes. J Theoret Biol. 2017 February;415:90-101.
63. Shenhav L, Zeevi D. Resource conservation manifests in the genetic code. Science. 2020;370(6517): 683–687.
64. Rozhoňová H, Payne JL. Little evidence the standard genetic code is optimized for resource conservation. Mol Biol Evol. 2021;38(11):5127-5133.
65. Xu H, Zhang J. Is the genetic code optimized for resource conservation? Mol Biol Evol. 2021b;38(11):5122-5126.
66. Massey SE. A neutral origin for error minimization in the genetic code. J Mol Evol. 2008;67(5):510-516.
67. Massey SE. The neutral emergence of error minimized genetic codes superior to the standard genetic code. J Theoret Biol. 2016 November;408:237-242.
68. Di Giulio M. A non-neutral origin for error minimization in the origin of the genetic code. J Mol Evol. 2018;86(9):593-597.
69. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324(5924):218-223. doi:10.1126/science.1168978
70. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, Tamir H, Achdout H, Stein D, Israeli O, et al. The coding capacity of SARS-CoV-2. Nature. 2021;589(7840):125-130. doi:10.1038/s41586-020-2739-1
71. Firth AE. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. Nucleic Acids Res. 2014;42(20):12425-12439.
72. Sealfon RS, Lin MF, Jungreis I, Wolf MY, Kellis M, Sabeti PC. FRESCo: finding regions of excess synonymous constraint in diverse viruses. Genome Biol. 2015;16(1):38. doi:10.1186/s13059-015-0603-7
73. Schlub TE, Buchmann JP, Holmes EC. A simple method to detect candidate overlapping genes in viruses using single genome sequences. Mol Biol Evol. 2018;35(10):2572-2581.
74. Nelson CW, Ardern Z, Wei X. OLGenie: Estimating natural selection to predict functional overlapping genes. Mol Biol Evol. 2020;37(8):2440-2449. doi:10.1093/molbev/msaa087
75. Louis AA. Contingency, convergence and hyper-astronomical numbers in biological evolution. Stud Hist Philos Biol Biomed Sci. 2016;58:107-116. doi:10.1016/j.shpsc.2015.12.014
76. Keefe AD, Szostak JW. Functional proteins from a random-sequence library. Nature. 2001;410(6829):715-718.
77. Çakır U, Gabed N, Brunet M, Roucou X, Kryvoruchko I. Mosaic translation hypothesis: Chimeric polypeptides produced via multiple ribosomal frameshifting as a basis for adaptability [published online ahead of print, 2021 Nov 7]. FEBS J. 2021;10.1111/febs.16269. doi:10.1111/febs.16269
78. Kosinski LJ, Masel J. Readthrough errors purge deleterious cryptic sequences, facilitating the birth of coding sequences. Mol Biol Evol. 2020;37(6):1761-1774.
79. Fernandes JD, Faust TB, Strauli NB, Smith C, Crosby DC, Nakamura RL, Hernandez RD, Frankel AD. Functional segregation of overlapping genes in HIV. Cell. 2016;167(7):1762-1773. e12. doi:10.1016/j.cell.2016.11.031
80. Safari M, Jayaraman B, Yang S, Smith C, Fernandes JD, Frankel AD. Functional and structural segregation of overlapping helices in HIV-1. Elife. 2022;11:e72482. doi:10.7554/eLife.72482
81. Dingle K, Ghaddar F, Šulc P, Louis AA. Phenotype bias determines how natural RNA structures occupy the morphospace of all possible shapes. Mol Biol Evol. 2022;39(1):msab280. Available from: https://doi.org/10.1093/molbev/msab280.
82. Schulz L, Sendker FL, Hochberg GKA. Non-adaptive complexity and biochemical function. Curr Opin Structur Biol. 2022 April;73:102339.
83. Gould SJ, Lewontin RC. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. Concept Iss Evol Biol. 1979;205:79.
84. Morris SC. Life's solution: Inevitable humans in a lonely universe. Cambridge: Cambridge University Press; 2003. Available from: https://doi.org/10.1017/CBO9780511535499