BioSystems 185 (2019) 104023

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

# Optimality in the standard genetic code is robust with respect to comparison code sets

Stefan Wichmann*, Zachary Ardern

*Department of Microbial Ecology, School of Life Sciences Weihenstephan, Technical University Munich, Weihenstephaner Berg 3, 85354 Freising, Germany*

ARTICLE INFO

*Keywords:*
Genetics
Genetic code
Evolutionary genetics

ABSTRACT

The genetic code and its evolution have been studied by many different approaches. One approach is to compare the properties of the standard genetic code (SGC) to theoretical alternative codes in order to determine how optimal it is and from this infer whether or not it is likely that it has undergone a selective evolutionary process. Many different properties have been studied in this way in the literature. Less focus has been put on the alternative code sets which are used as a comparison to the standard code. Each implicitly represents an evolutionary hypothesis and the sets used differ greatly across the literature. Here we determine the influence of the comparison set on the results of the optimality calculation by using codes based upon different sub-structures of the SGC. With these results we can generalize the results to different evolutionary hypotheses. We find that the SGC's optimality is very robust, as no code set with no optimised properties is found. We therefore conclude that the optimality of the SGC is a robust feature across all evolutionary hypotheses. Our results provide important information for any future studies on the evolution of the standard genetic code. We also studied properties of the SGC concerning overlapping genes, which have recently been found to be more widespread than often believed. Although our results are not conclusive yet we find additional intriguing structures in the SGC that need explanation.

## 1. Introduction

One nearly universal feature of life as we know it is the standard genetic code (SGC). While alternative genetic codes do exist, all extant codes are probably derivatives of the SGC, as they are taxonomically restricted and therefore considered younger (Knight et al., 2001) and they differ by few codon assignments. In this study we assume that the SGC has always been a true code as it is today, which means that the mapping is arbitrary i.e. any codon could map to any amino acid (Barbieri, 2018). This can be achieved by adapter molecules such as tRNAs that can map any amino acid to any codon. We assume that a tRNA-like molecule completely defined the code at some point, since all amino acids that have shown a tendency towards stereochemical binding to their codons or anticodons thus far (Yarus et al., 2009) are not prebiotically available and have high energy costs of synthesis (Higgs and Pudritz, 2009). Even if the code was stereochemically determined in the beginning, we cannot see any traces of this left in the present SGC. Without stereochemical binding determining the code, hypotheses about whether the SGC is an adaptation are possible, along with investigating which properties are adaptive. The best-studied property is the robustness to mutational or misread errors in genes (Freeland and Hurst, 1998b; Haig and Hurst, 1991). Other properties of the code which have been hypothesized to be near optimal include at least five properties: minimizing the toll of frameshift errors (Itzkovitz and Alon, 2007), supporting finding functional nucleotide sequences by random mutations (Tripathi and Deem, 2018), facilitating coding of additional information alongside protein sequences (Itzkovitz and Alon, 2007), conserving sequences in alternative reading frames (Konecny et al., 1993) and creating alternating hydropathy patterns in sense–antisense protein pairs (Blalock, 1990; Zull and Smith, 1990).

There are two main kinds of approach to studying the optimality of the SGC, aiming at different aspects. The older approach is the 'statistical approach', which compares the SGC to a set of theoretical alternative codes and determines the percentage of better codes in a chosen property. This approach was developed in order to test the null hypothesis that the SGC has not been optimized and is a 'frozen accident' (Crick, 1968) randomly picked from the set of possible codes. In the evolution of the SGC have most likely been factors that restrict what the code can look like and therefore determine the set of possible codes. After proposing an evolutionary hypothesis we can construct this set of possible codes and determine the percentage of codes better than the SGC. Following Massey (2008) if the chance to draw a code better than

* Corresponding author.
*E-mail address:* stefan.wichmann@tum.de (S. Wichmann).

https://doi.org/10.1016/j.biosystems.2019.104023
Received 15 May 2019; Received in revised form 22 August 2019; Accepted 24 August 2019
Available online 11 September 2019

the SGC is below 5% we can reject the null hypothesis of no optimization, and infer a likely optimization process.

A more recent approach to code optimality is an 'engineering approach', which uses a genetic algorithm (GA) to search for the best and the worst code for a property following a chosen evolutionary hypothesis. The GA starts with a random code, which is run through multiple cycles of variation and selection in order to find the extreme values. Since GAs can be stuck in local extrema, the GA is run with different random starting codes. Finding the best and worst codes allows calculation of the distances of the SGC to these codes (Błażj et al., 2018) and also allows inferring the scale of optimization and how difficult it is to evolve the code. But it is not possible to determine whether the SGC has been optimized at all in this approach. If the set of possible codes is not distributed symmetrically around some mean value, it could be that the worst code is much further away from most codes than the best code is. So, even if the distance of the SGC to the best code is much smaller than to the worst code it does not necessarily mean that the SGC has been optimized. Also it is not clear what level of optimization for a given property we should expect under a hypothesis of selection. Assuming the SGC has evolved through selection, the SGC could have evolved to a local or even global fitness peak, but the fitness of the SGC may not only be determined by a single property but simultaneously by many. Most studies only consider a single property as it is not clear how to weight different properties and combine them into a fitness function. Studying only a single property might create lower values of optimality. It is also not clear how far the SGC has been optimized. The process could have been stopped before a maximum was reached. One possible cause is Crick's concept of the 'frozen accident' (Crick, 1968), namely that the organism carrying the SGC was complex enough such that a change in the SGC would be more deleterious than beneficial. Another possible scenario is fixation of the code following the evolution of horizontal gene transfer (HGT). In order for HGT to be efficient all organisms need to have the same genetic code, which has been shown to lead to stable convergence to a single code (Aggarwal et al., 2016). An objection to the optimization hypothesis is that it appears that many steps are needed on average to create a code as optimal as the SGC (Massey, 2010). However this does not make the optimization hypothesis hugely more unlikely than any alternatives, as the evolution of the SGC was noted as a 'notoriously difficult problem' from the beginning (Crick et al., 1976), before the optimization was recognised, and remains so today (Kun and Radványi, 2018).

In this study we are interested in determining whether the genetic code shows evidence of having been optimized or not, thus we use the 'statistical approach'. We will not discuss the different hypothesis on how the SGC evolved, but refer the interested reader to a recent review (Kun and Radványi, 2018). Previous studies show that results of the 'statistical approach' heavily depend on the codes used as a comparison to the SGC. For the mutational error robustness the percentage of better codes than the SGC ranges from 0.0001% (Freeland and Hurst, 1998b) to 21.9% (Massey, 2008), depending on which evolutionary hypothesis is used. In order to remove the optimality for the mutational robustness a very specific evolutionary hypothesis had to be used in (Massey, 2008). They show that after relaxing the constraints of the hypothesis only a little the mutational error robustness returns to optimal in the SGC, so their result is not very robust. Assuming that their evolutionary pathway is not exactly correct, the claim that the mutational error robustness of the SGC really had a neutral origin is questionable in itself, though our analysis adds to this.

We test multiple properties on different code sets in order to infer the robustness of the optimality of the SGC. If the genetic evolved via natural selection, it should yield a fitness advantage for some kind of replicating RNA/DNA system otherwise it would be expected to be lost (Kun and Radványi, 2018). Since the most basic function of the genetic code is to translate from mRNA to proteins, we assume that the created proteins have a function and that the mRNA has to be translated mostly without errors. In order to do so in an energy efficient and reliable way, the genetic code should be robust against mutational and misread errors, which is the first property we test. A second type of translation error is a frameshift error, which is a shift of the ribosome by one or two nucleotides on the mRNA. This results in a completely different protein being translated and should be stopped as soon as possible, since protein translation is energetically costly (Lynch and Marinov, 2015). The same effect can be observed when an insertion or deletion of nucleotides takes place. In order to reduce the fitness cost of such a frameshift event, the genetic code could be designed in such a way that the other two sense reading frames of the translated strand have a high STOP codon probability. This would stop the faulty translation shortly after the frameshift and can be achieved by pairing the codons most frequently found in the other two frames with STOP codons when frameshifted (Itzkovitz and Alon, 2007). Just as in Itzkovitz and Alon (2007) and Mir and Schober (2014) we study the mean value of both sense reading frames as both should be optimized at the same time, but also both frames individually.

Two more properties we study only matter for overlapping genes (OLGs). A genome containing OLGs will be shorter and therefore increase replication rates as the genome has to be copied in any replicating system. In a competitive setup this would plausibly yield a fitness advantage and could be selected for. Viruses are known to carry multiple OLGs (Barrell et al., 1976; Fiddes and Godson, 1978), but these fascinating genes are also found in prokaryotes (Fukuda et al., 1999; Tunca et al., 2009; Kim et al., 2009; Hücker et al., 2018b,a; Vanderhaeghen et al., 2018), eukaryotes (Spencer et al., 1986; Iwabe and Miyata, 2001) and even vertebrates (Williams and Fried, 1986; Makalowska et al., 2004). It has intriguingly been proposed that amino acyl tRNA synthetases arose from a sense-antisense overlapping gene pair (Rodin and Ohno, 1995; Martinez-Rodriguez et al., 2005). In theory, creating OLGs is difficult since the 'mother gene' restricts the alternative reading frames, which makes it difficult to encode functional proteins. Even if a functional OLG is created, any mutation in the overlapping region will potentially damage two genes, so the OLG is more likely to be lost in the course of evolution than a regular gene. Due to the difficulties in creating and maintaining OLGs they have long been thought to only appear in viruses, where genome size is an important factor. Nevertheless OLGs exist outside of the virus domain, so an explanation of how an organism can overcome these difficulties is needed. One plausible route of explanation is that the structure of the genetic code facilitates the creation and/or maintenance of OLGs. A related question regarding the susceptibility of OLGs to mutations has been studied by Konecny et al. (1993): if we have a conservative nucleotide change in the '+1' frame, which is a change that does not change the amino acid coded for by the codon, how would this affect the '-1' frame? To partly counter the mutational vulnerability of OLGs the genetic code could conceivably be optimized to be especially error resistant in alternative reading frames when there is a conservative mutation in the '+1' frame. This property can be optimised in a certain reading frame if all codons of a single amino acid overlap with codons of similar amino acids in that frame. Here we expand the study done in Konecny et al. (1993) on the '-1' frame to all alternative reading frames, using the reading frame definitions summed up in Fig. 1.

In order for OLGs to be able to code for all kinds of proteins, the average open reading frame (ORF) length on alternative reading frames should ideally be similar to non-OLGs. The average ORF length is a good measure for whether any of the alternative reading frames stands out from the others. If OLGs were an important feature for organisms at the time of genetic code optimization, a high average ORF length may have yielded an evolutionary advantage. This property is optimised if STOP codons share nucleotides with rarely used codons, i.e. STOP codons overlap rarely used codons in alternative reading frames. In each frame, this property conflicts strongly with reduction of the impact of frameshift errors, especially for reading frames on the same strand, and is the last property we study. But different reading frames could have different functions. Some could be adjusted for long genes, while others

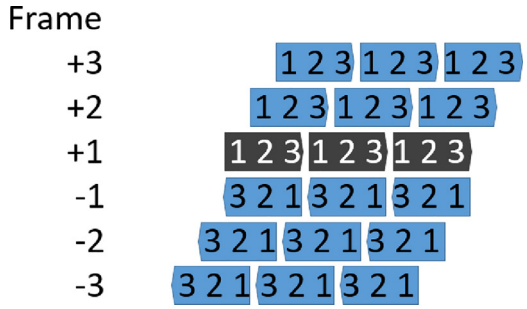


**Fig. 1.** Illustration of the alternative reading frame definitions.

quickly stop the translation after a frameshift error. In this sense both properties can be realized in one genetic code – but whether different reading frames are actually used differently in accordance with their properties deserves further exploration.

Each code set we compare the SGC with in any of these four properties implicitly represents an evolutionary hypothesis. The SGC has a few unusual structures compared to a completely random code, for example only one out of $10^{65}$ random codes has the same block structure as the SGC. Since the SGC has such rare structures, all but the stereochemical hypothesis of code evolution (Woese et al., 1966) try to explain the structures of the standard genetic code. So instead of using existing evolutionary hypothesis to create theoretical alternative code sets, we will use codes created from different structures of the SGC. By comparing the structures of the SGC that an existing or new evolutionary hypothesis enforces on the set of possible codes we can infer the order of optimality from our analysis. This allows us to make more general claims about the robustness of the optimality of the SGC.

We study three different kinds of structures of the SGC and completely random codes called the 'Random' code set as comparison. The first type of structure is the composition of the SGC, which is determined by the number of STOP codons, the number of different amino acids and the number of codons coding for each amino acid. The respective code sets will be labelled 'Random_fs' (fixed STOPs), 'Random_faa' (fixed amino acids) and 'Degeneracy'. Next we study the absolute structure in the SGC, which is the degeneracy of the third codon position. We call this the 'Blocks' code set as it conserves the blocks of codons coding for the same amino acid. Lastly we study relative structures of the SGC, which is the similarity of amino acids with codons which differ only in one nucleotide. We create similar neighbour code sets from the 'Degeneracy' code set called 'Degeneracy_n' and from the 'Blocks' code set called 'Blocks_n'. We also use the code set created in (Massey, 2008) using the 2-1-3-model of code evolution (Massey and Sequential, 2006). In this hypothesis the block structure of the SGC is conserved, and additionally, neighbouring blocks have similar amino acids created by a specific scheme originating from a sequential introduction of more and more amino acids to an originally simple code consisting only of valine, alanine, aspartic acid and glycine. This is the only comparison code set so far that results in no optimality for the SGC in mutational error robustness. We also tested a few more code sets, namely a generalized way of creating random blocks taken from Buhrman et al. (2011), a random code set with both the number of STOPS and different amino acids fixed, and two versions of the same sets with similar neighbours; the data for these can be found in Figs. S2–S6 of the Supplementary Material.

## 2. Results and discussion

### 2.1. Mutational and misread error robustness

The optimality of mutational robustness is a very robust feature of the SGC, as the percentage of better codes is far below the 5% threshold for all code sets except the 213-model, which is slightly above the threshold, c.f. the left panel of Fig. 2. For the 'Degeneracy' and the 'Degeneracy_n' code set not a single code better than the standard code was found in $10^{10}$ codes.

Neither the STOP codons nor the number of different amino acids in the code have a strong influence, c.f. the right panel of Fig. 2. Fixing the degeneracy or introducing the block structure has a similar effect on the mutational robustness on average. Since the block structure also creates very similar amino acid degeneracies as in the SGC this feature seems to fix the average value. Fixing the degeneracy on the third nucleotide position increases the variance as either very similar amino acids are close or very different ones. Creating codes with similar neighbours barely influences the optimality, and only the specific evolutionary path in 213-Model can, so the details of the evolutionary hypothesis matter a great deal. The exact percentage of better codes for all code sets can be found in Table S1 in the Supplementary Material.

### 2.2. Frameshift error abortion times

The distance to the next STOP codon after frameshift is only optimal in the 'Degeneracy' and the 'Degeneracy_n' code set when considered in each sense reading frame independently, but the average of both frames is also optimal in the 'Random_fs' and the '213_Model' code set, c.f. top panel of Fig. 3. Optimizing both reading frames at the same time is harder than optimizing only a single reading frame but more relevant in a realistic scenario, so it is the property that should be studied instead

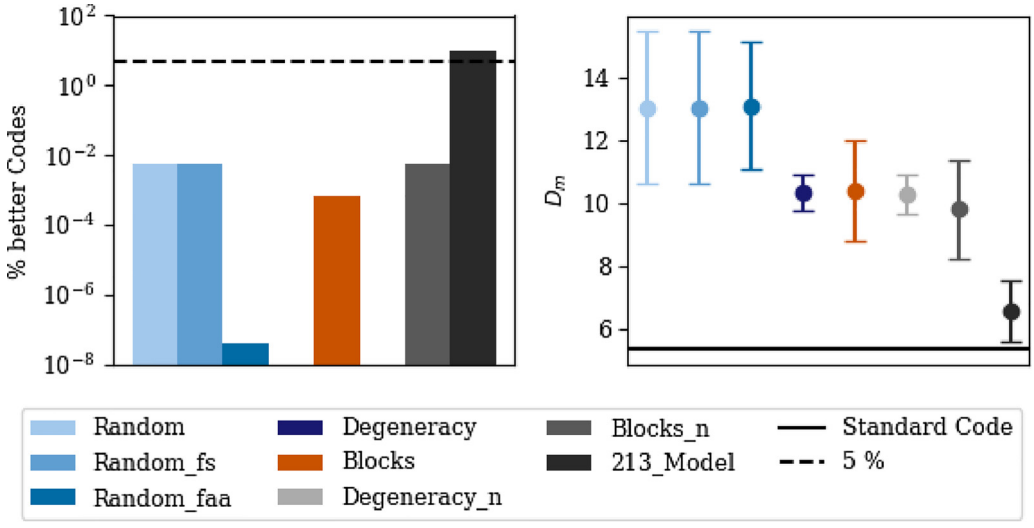


**Fig. 2.** Mutational and misread error robustness calculated on $10^{10}$ codes. Compositional code sets are shown in different shades of blue, absolute structural code sets are shown in different shades of orange and relative structural code sets are shown in different shades of grey. *Left:* Optimality as the percentage of better codes. The black line indicates the 5% threshold. *Right:* Mean values with standard deviation as error bars. The black line indicates the $D_m$ value for the standard code.

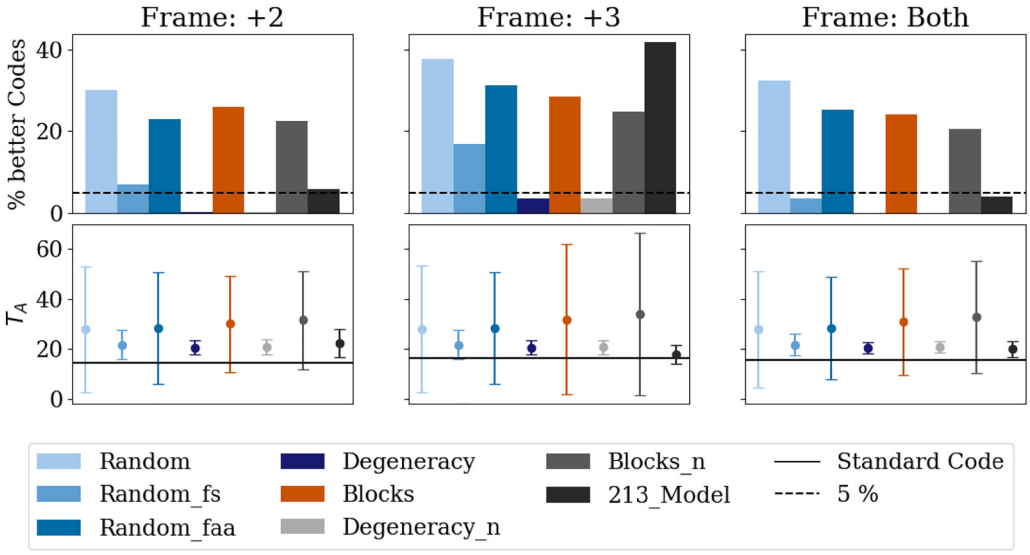


**Fig. 3.** Frameshift error abortion times calculated for $10^5$ codes. Compositional code sets are shown in different shades of blue, absolute structural code sets are shown in different shades of orange and relative structural code sets are shown in different shades of grey. *Top:* Optimality as the percentage of better codes. The black line indicates the 5% threshold. *Bottom:* Mean values with standard deviation as error bars. The black line indicates the $T_A$ value for the standard code.

of single reading frames. The previously mentioned code sets all restrict the number of STOP codons to 3 and thus have very small variations across their code set values leading to a higher optimality of the SGC, c.f. bottom panel of Fig. 3. The distance to a STOP codon is bounded below by 0, while there is no upper bound for a code without a STOP codon. We expect that the more STOP codons are included in the SGC the less impact each additional STOP codon has on the result. This could be one factor determining that only 3 STOP codons are included in the SGC.

The other structures besides the number of STOP codons have little impact on this property. Interestingly, we find the distance to a STOP codon after a frameshift to be optimal in the '213_Model', so removing the mutational error robustness with the underlying evolutionary hypothesis does not remove the optimality of the SGC. In this study we only test four properties, but additional optimalities are possible considering the multiple properties of the SGC. The exact percentage of better codes for all code sets in the sense reading frames can be found in Table S1. The anti-sense strand could also have some absorbing reading frames, but it is not yet clear which reading frames to consider for coding an which for absorbing faulty translations. The data of each anti-sense reading frame can be found in Fig. S4 and Table S2 in the Supplementary Material.

### 2.3. Conservation in alternative reading frames

The absolute conservation ($D_c$) values of the '-1' frame cannot be compared to other reading frames, as in this frame a single codon on the '+1' frame defines a full codon, while on all other frames two codons in the '+1' frame are needed. This leads to 20 groups, one for each amino acid, on which this property is calculated in the '-1' frame, but in all other frames 400 groups, one for each dipeptide, are used. The values of the '-1' frame are roughly 20 times larger than the values in other alternative reading frames and we expect this to be an artifact of calculation, but have not studied this in more detail, c.f. bottom panel of Fig. 4.

Only when compared to the 'Degeneracy' and the 'Degeneracy_n' code set is the SGC optimal across all reading frames, c.f. top panel of Fig. 4. Again we find optimality in the '213_Model' code set, namely in the two sense reading frmes and the '-3' frame. In the '-1' frame almost all code sets are optimal except the 'Blocks', the 'Blocks_n' and the '213_Model'. If any reading frame has been optimised to be especially conservative it is the '-1' frame, but the relative optimality still depends on the evolutionary background.

No structure of the SGC strongly influences the conservation value except in the case of the '-2' frame. Since the third codon positions in the '+1' and the '-2' overlap, codes with the degeneracy on the third codon position are especially conserved, as synonymous mutations in the mother gene are also mostly synonymous in the '-2' frame. Interestingly, the SGC, which also has this degeneracy structure on the third codon position, is not influenced on the '-2' frame. Extending this insight, when comparing the conservation values of the SGC across reading frames except the '-1' frame, we find them to be remarkably similar. Just as for the frameshift error reduction property, studying single reading frames might not be the right approach for this property. In the case of OLGs we do not only want to conserve existing genes but plausibly also need some coding flexibility in order to create an OLG to begin with. Since conservation and flexibility are opposing properties, a trade-off between the two properties might lead to a constant value across reading frames being optimal. Further studies are needed in order to clarify this intriguing result. The exact percentage of better codes for all code sets can be found in Table S2 in the Supplementary Material.

### 2.4. Average ORF length

Only in the '-2' frame do we find optimality of the SGC. The SGC is optimal with regards to the 'Degeneracy', the 'Degeneracy_n' and the '213_Model' code sets for this reading frame, c.f. top panel of Fig. Fig. 5. The latter model gives especially high optimality as only 0.012% of codes are better than the standard code. As the average ORF length and the distance to a STOP codon after a frameshift error are very similar properties, the influence of the number of STOP codons is the same in both properties, c.f. bottom panel of Fig. 5. Besides the number of STOP codons we do not see a clear influence of any other structure on the average ORF length, except in the '-2' frame, where the degeneracy structure on the third nucleotide position greatly increases the variability among theoretical alternative codes. The reason for this influence is not clear, but since the two third codon positions overlap in this reading frame, and this position is especially important for this structure, we expect a connection between the two observations. The exact percentage of better codes for all code sets can be found in Table S3 in the Supplementary Material.

## 3. Conclusion

It has been shown in the literature that the underlying evolutionary hypothesis used to create theoretical alternative codes can strongly influence the results of an optimality calculation, but we find that the

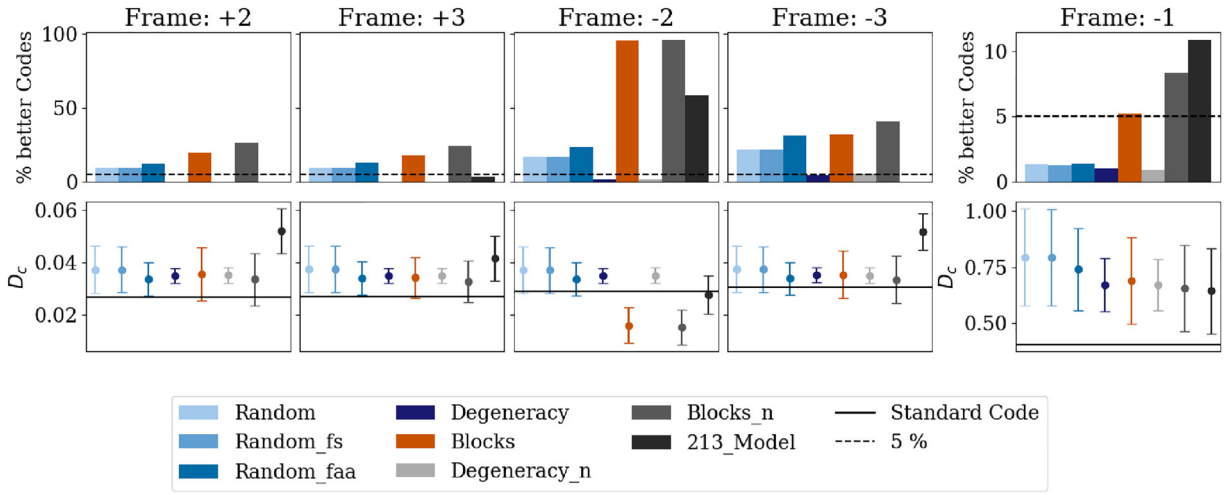


**Fig. 4.** Conservation in alternative reading frames calculated for $10^7$ codes. Compositional code sets are shown in different shades of blue, absolute structural code sets are shown in different shades of orange and relative structural code sets are shown in different shades of grey. *Top:* Optimality as the percentage of better codes. The black line indicates the 5% threshold. *Bottom:* Mean values with standard deviation as error bars. The black line indicates the $D_c$ value for the standard code.

optimality of the SGC is still a very robust feature, when considered as a 'multi-dimensional' property. In all code sets but the '213_Model' we find the SGC to be very optimal in the mutational robustness. Not even our algorithm to create code sets with similar neighbours nor those in the first two models tested in Massey (2008) are able to produce many codes as good as the SGC. Only by choosing a very specific evolutionary hypothesis as in the '213_Model' can the mutational robustness be explained without an optimization step in the SGC. But even if the mutational robustness is not optimal in the '213_Model', various other SGC properties turn out to be optimal in this code set. Some properties concern OLGs, but also the distance to a STOP codon after a frameshift error is found to be optimal in the '213_Model'; and we have only tested a few of the possible properties of the SGC. Therefore we conclude that the optimality of the SGC is a robust feature and cannot be explained by any simple evolutionary hypothesis proposed so far.

After testing many properties for optimality we can expect that a few will turn out optimal just by chance. Therefore it is very important when fitting these results into an evolutionary context that the tested properties plausibly create a fitness advantage for the replicating system. Quantifying a threshold above which an effect could be selected for is very difficult however. The effect strength of the fitness advantage must overcome stochastic fluctuations in the population, which depend on population size. It is not yet possible to estimate this threshold for early life.

In this study we tested each property individually, but we saw in the distance to a STOP codon after a frameshift error that only a combination of both reading frames turned out to be optimal, which in any case is the most sensible property as long as there are no frequency differences in frame shift errors between the '+2' and the '+3' frame. Taking this observation further, in the ideal case all properties that the SGC is supposedly optimal in should not be tested individually but combined into a single fitness function in order to mimic a real selection process. This is very difficult as it is not clear what fitness contribution each property adds to the overall fitness. A simpler but less realistic approach that comes to mind is to test properties in a sequential manner. i.e. collect all codes that are better than the SGC in one property and test the next property on this subset of codes. Repeating this process for all properties we can then find conditional optimalities. Something similar has been done in Itzkovitz and Alon (2007) and Mir and Schober (2014), which both use a code set constructed in such a way that the mutational robustness of the SGC is conserved. Unfortunately this only works when all properties are independent of each

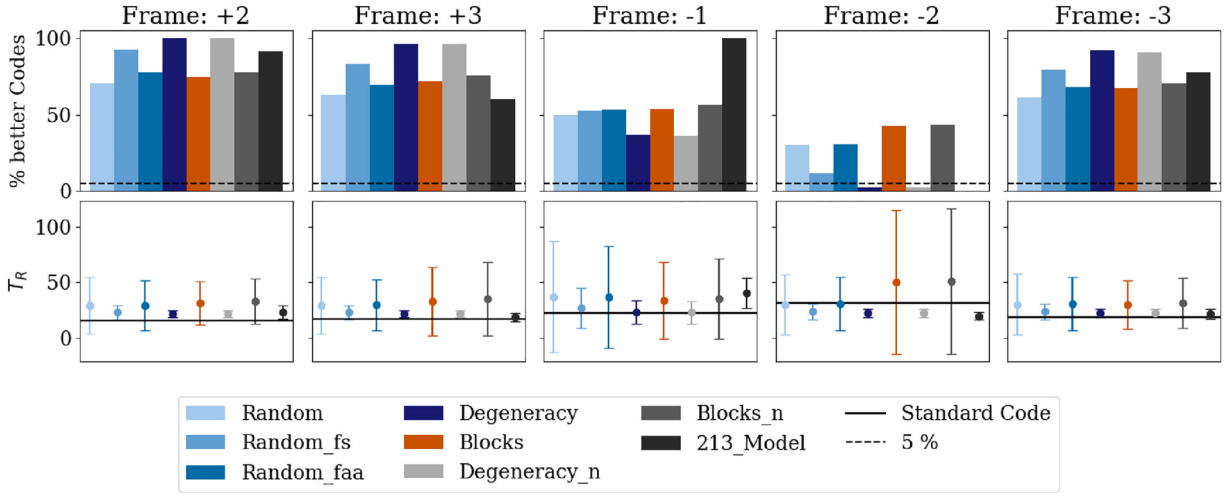


**Fig. 5.** Average ORF length calculated for $10^5$ codes. Compositional code sets are shown in different shades of blue, absolute structural code sets are shown in different shades of orange and relative structural code sets are shown in different shades of grey. *Top:* Optimality as the percentage of better codes. The black line indicates the 5% threshold. *Bottom:* Mean values with standard deviation as error bars. The black line indicates the $T_R$ value for the standard code.

other, otherwise the order in which the properties are being tested will influence the result; and there is no natural ordering of the properties. For example, the distance to a STOP codon after a frameshift error and the average ORF length are almost the same property, but the first is optimal for short distances to a STOP codon, while the latter is optimal for long distances. If all properties were independent of each other, testing them individually would be sufficient as it yields the same results. Only combining all properties to a single fitness function could finally answer the question of code optimality within a framework where optimality is taken to imply both an optimization process and natural selection as the driving force of optimization.

In the literature and also in our study many properties have very high reported optimalities, meaning that the probability of finding the standard genetic code by chance is very low. Selection is not an omnipotent force, so this raises the question of whether a selection process could have found the SGC in the case of extreme code optimalities. For some evolutionary hypotheses this has already been tried (Massey, 2010), but this question is strongly related to how many codes could have been tested by natural selection during code evolution, which is not yet answered. By studying these questions we might be able to rule out some evolutionary hypotheses and thereby further our understanding of the evolution of the SGC. Conducting a similar study to this with many different evolutionary hypotheses, but using GAs to determine how many codes must be tested in order to find genetic codes similar to the SGC, would greatly extend our knowledge on this topic.

We believe of particular interest for future research is that we tested just two properties connected to OLGs. The conservation is mostly only optimal in the '-1' frame, while the average ORF length in only optimal in the '-2' frame. Taking just these results, OLGs do not seem to be clearly optimized for overall, but on closer inspection there are many questions left. The average ORF length was calculated as the average length between STOP codons, but some STOP codons can be removed by synonymous mutations in the mother gene. It might be more realistic to study the average distance between two STOP codons that cannot be removed by judicious codon usage. In the conservation of OLGs we saw that the SGC has unexpectedly similar values across all readying frames (except the '-1'). It is possible that this value represents an optimum for this property as existing OLGs not only need to be preserved but new OLGs must be created, so coding flexibility is also needed. A low optimality in a value does not necessarily mean that the property has not been optimised for or that it is not important but maybe the optimum has already been reached due to other factors. In this case it could be a trade-off between conservation and flexibility. In this study we only tested two properties, but more are of interest concerning OLGs. Just to name one, the 'mother gene' could enforce structure on alternative reading frames via the genetic code, making formation of *de novo* genes more likely.

The idea of trade-offs between properties of the standard genetic code could be taken further. For example, the number of different amino acids in the SGC could face a trade-off. Having more amino acids opens up a wider protein space, but also reduces the mutational robustness as the degeneracy structure strongly improves this property. We propose that the idea of trade-offs could be very important in understanding the nature of SGC optimality. Every property has a cost versus some other property and assuming that the genetic code had some freedom in its evolution, some trade-offs were plausibly experienced as constraints. Finding further such trade-offs may launch future studies into this central topic in molecular biology which surprisingly remains unexhausted.

## 4. Materials and methods

### 4.1. The random code set

This subset has no restrictions besides the set of 20 amino acids or a STOP codon, which the codes can include, although not all 20 amino acids or a STOP codon have to be included. It embodies the total possibility space for triplet genetic codes with the same total pool of possible products as the SGC, and every other code set will be a subset of this 'Random' set.

### 4.2. Composition code sets

The 'Random_fs' (fixed STOPS) code set restricts the number of STOP codons to exactly three, the 'Random_faa' (fixed amino acids) code set only contains random codes which include all 20 amino acids and the 'Random_fb' (fixed both) code set is the intersection of the first two sets, so it contains only codes with all 20 amino acids and exactly three STOP codons. The 'Degeneracy' code set is created by restricting each amino acid in its codes to exactly as many codons as the SGC uses.

### 4.3. Absolute structure code sets

We construct the 'Blocks' code set by combining all codons which code for the same amino acid in the SGC into a fixed 'block' and changing the amino acid assigned to each block. In our calculations this also includes the STOP codons, which in the literature are often left fixed. This relaxes both the amino acid degeneracy and the number of STOP codons, but only small variations are possible.

A more randomized version of the 'Blocks' code set was constructed in Buhrman et al. (2011). They used the wobble binding rules on the last nucleotide (Agris et al., 2007; Berg et al., 2010) to construct all possible boxes of degeneracies on the last nucleotide, where a box contains all codons which have the same first two nucleotides. Drawing random boxes and fixing the number of STOP codons to three and the number of amino acids to 20 we can study the influence of the degeneracy on the third codon position in a more general approach. We call this code set the 'Random_Blocks' code set.

### 4.4. Relative structure code sets

The degeneracy on the third nucleotide position already creates similar neighbours, but in addition to this the blocks of amino acids can be arranged in order to make neighbours as similar as possible. These code sets are created by first picking a random codon/block from the code and listing the amino acids on the neighbouring codons/blocks. Next we create a list of all remaining amino acids to be incorporated in the code which have a smaller distance than a threshold distance $d$ to at least one of the neighbouring amino acids. If this list is not empty we choose a random amino acid from the list. Otherwise a random amino acid from all remaining amino acids is selected. We will use polar requirement (Mathew and Luthey-Schulten, 2008) differences to determine similarity between amino acids. We found that a threshold distance of one fourth of the standard deviation of the polar requirement values of all 20 amino acids in the SGC creates codes with the most similar neighbours. This can be done for every code set but we only do it for the 'Random_fb', the 'Degeneracy', the 'Random_Blocks' and the 'Blocks' code set resulting in the 'Random_fb_n', the 'Degeneracy_n', the 'Random_Blocks_n' and the 'Blocks_n' code set.

### 4.5. Mutational and misread error robustness

It is known that the standard code has a high mutational robustness (Haig and Hurst, 1991), a finding which is highly robust to changes in calculation details. Both misreads and mutations have been shown to appear with different frequencies for transitions and transversions (Collins and Jukes, 1994; Kumar, 1996; Moriyama and Powell, 1997; Morton, 1995; Friedman and Weinstein, 1964) but the observed differences vary strongly and most likely depend on many details of calculation or experiment. Also a codon is not read with the same accuracy on all positions, namely the second position has the highest accuracy, followed by the first position, while the third position is the most error

prone (Parker, 1989; Woese, 1965). Introducing different weights for mutations or misreads on different positions and for transitions and transversions increases the optimality (Freeland and Hurst, 1998b,a; Freeland et al., 2000), but the values of the weights used in these studies represent a tendency rather than a quantitative value, so we will refrain from using weights in this study. The measure for the mutational robustness is the average change a single mutation inflicts on a gene. In order to calculate this property a numerical distance between different amino acids has to be defined. Here we will use polar requirement (Woese et al., 1966; Spencer et al., 1986) to define the distance function. Since distances to STOP codons cannot be defined that way we will use a suppression approach for STOP codons as has been suggested (Buhrman et al., 2011). The suppression approach does indeed minimize the effect of STOP codons (Fig. S1 in the Supplementary Material).

Following Buhrman et al. (2011), Eq. (1) is the formula of calculation for the MMER $D_m$, where $d(a_i, a_j)$ is the difference of polar requirement values of the amino acids $a_i$ and $a_j$ and $n_{STOPs}$ is the number of STOP codons in each code.

$$D_m = \frac{\sum_{i,j} d^2(a_i, a_j)}{9(64 - n_{\text{STOPs}})} \tag{1}$$

### 4.6. Frameshift error abortion time

Following Mir and Schober (2014), we will calculate the average number of codons translated after a frameshift before a STOP codon is encountered. Following Mir and Schober (2014), Eqs. (2)–(4) define the average number of amino acids before a STOP codon $T_A$, where $P(c_j|c_i)$ is the conditional probability that codon $c_j$ follows codon $c_i$.

$$T_A = \frac{\sum_i t_f^{(i)}}{64} \tag{2}$$

$$\vec{t_f} = (\hat{1} - \hat{Q}_f)^{-1} \cdot \vec{1} \tag{3}$$

$$[\hat{Q}_f]_{i,j} = P(c_j|c_i) \tag{4}$$

### 4.7. Conservation in alternative reading frames

First we construct the sets of codons produced by conservative mutations, while also taking two and three point mutations into account so long as the result is conservative (i.e. results in no change in the amino acid). In the '-1' frame this is straightforward, as we only have to translate all codons of each amino acid into the '-1' frame and thus create sets of codons. Picking two codons $i$ and $j$ from such a set and calculating the distance between their respective amino acids $d(a_i, a_j)$ we can estimate the difference resulting from conservative mutations. Just as for the mutational robustness we use the squared amino acid distance. In Eq. (6) $n_i$ is the number of amino acids coding for amino acid $a_i$ and $P_i$ the probability to find this amino acid in a gene.

$$\bar{d}_a = \frac{2}{n_a(n_a - 1)} \sum_{i,j,i\neq j} d^2(a_i, a_j) \tag{5}$$

$$D_c = \sum_a \frac{n_a - 1}{\sum_b (n_b - 1)} P_a \bar{d}_a \tag{6}$$

For all other alternative reading frames we have to construct groups for all possible combinations of two amino acids. From each combination of amino acids we extract the codon in each alternative reading frame. In order to reduce double counting, only conservative mutations that change the nucleotides of the codon on the alternative reading frame will contribute to the set of codons produced by conservative mutations. The only difference to the '-1' frame case is that the sum over a in Eq. (6) runs over all sequences of dipeptides and $P_a$ are the occurrence probabilities of those sequences. Details of derivation of equations (5) and (6) as well as examples to calculate (5) can be found in the Supplementary Material.

### 4.8. Average open reading frame (ORF) length

We use the same approach as Mir and Schober (2014), who has estimated the average open reading frame length as the average number of codons between two STOP codons (Sieber et al., 2018). Following Mir and Schober (2014), Eq. (7) defines the calculation of the average ORF length $T_R$, where $P(c_{\text{STOP}}^{(i)})$ is the probability to encounter the STOP codon $c^{(i)}$.

$$T_R = \frac{1}{P_{\text{STOP}}} = \frac{1}{\sum_i P(c_{\text{STOP}}^{(i)})} \tag{7}$$

### 4.9. Codon probabilities

In some of the properties codon and conditional probabilities are used. These probabilities can be extracted from any genome; in this study we use the pathogenic bacterium Escherichia coli O157:H7 EDL933 (Accession number NC 002655, EHEC).

The codon usage statistic could have adapted to the standard code such that using this statistic for alternative codes would cause the standard code to artificially appear more optimal. Therefore, just as in Mir and Schober (2014) we will not use codon statistics but amino acid statistics. Amino acids make up the proteins and are therefore the building blocks of life. Their usage statistic depends on what is needed to make the necessary proteins. Codon statistics can be constructed from the amino acid statistics by assuming each codon coding for the same amino acid has the same probability. The probabilities in the '+1' frame are determined by amino acid usage statistics. The conditional probabilities in the other reading frames can be calculated from the '+1' frame as shown in Mir and Schober (2014).

## Conflict of interest

None declared.

## Acknowledgement

This work was supportet by Freistaat Bayern through the Technical Univeristy of Munich.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.biosystems.2019.104023.

## References

Aggarwal, N., Bandhu, A.V., Sengupta, S., 2016. Finite population analysis of the effect of horizontal gene transfer on the origin of an universal and optimal genetic code. Physical Biology 13. https://doi.org/10.1088/1478-3975/13/3/036007.

Agris, P.F., Vendeix, F.A.P., Graham, W.D., 2007. tRNA's wobble decoding of the genome: 40 years of modification. Journal of Molecular Biology 366 (1), 1–13. https://doi.org/10.1016/j.jmb.200611.046.

Błażj, P., Wnetrzak, M., Mackiewicz, D., Mackiewicz, P., 2018. Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm. PLOS ONE 13 (8). https://doi.org/10.1371/journal.pone.0201715.

Barbieri, M., 2018. What is code biology? Biosystems 164, 1–10. https://doi.org/10.1016/j.biosystems.2017.10.005.

Barrell, B.G., Air, G.M., Hutchison III, C.A., 1976. Overlapping genes in bacteriophage φx174. Nature 264, 34–41. https://doi.org/10.1038/342640a0.

Berg, J.M., Tymoczko, J.L., Stryer, L., 2010. Biochemistry, 7th ed. W.H. Freeman and Company, New York.

Blalock, J.E., 1990. Complementarity of peptides specified by 'sense' and 'antisense' strands of DNA. Trends in Biotechnology 8, 140–144. https://doi.org/10.1016/0167-7799(90)90159-UF.

Buhrman, H., van der Gulik, P.T.S., Kelk, S.M., Koolen, W.M., Stougie, L., 2011. Some mathematical refinements concerning error minimization in the genetic code. IEEE/

ACM Transactions on Computational Biology and Bioinformatics 8 (5), 1358–1372. https://doi.org/10.1109/TCBB.201140.
Collins, D.W., Jukes, T.H., 1994. Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics 20 (3), 386–396. https://doi.org/10.1006/geno.1994.1192.
Crick, F.H.C., Brenner, S., Klug, A., Pieczenik, G., 1976. A speculation on the origin of protein synthesis. Origins of life 7 (4), 389–397. https://doi.org/10.1007/BF00927934.
Crick, F.H.C., 1968. The origin of the genetic code. Journal of Molecular Biology 38 (3), 367–379. https://doi.org/10.1016/0022-2836(68)90392-6.
Fiddes, J.C., Godson, G.N., 1978. Nucleotide sequence of the J gene and surrounding untranslated regions of phage G4 DNA: Comparison with phage φx174. Cell 15 (3), 1045–1053. https://doi.org/10.1016/0092-8674(78)90288-X.
Freeland, S.J., Hurst, L.D., 1998a. Load minimization of the genetic code: history does not explain the pattern. Proceedings of the Royal Society B: Biological Sciences 265 (1410), 1229–2111. https://doi.org/10.1098/rspb.1998.0547.
Freeland, S.J., Hurst, L.D., 1998b. The genetic code is one in a million. Journal of Molecular Evolution 47 (3), 238–248. https://doi.org/10.1007/PL00006381.
Freeland, S.J., Knight, R.D., Landweber, L.F., Hurst, L.D., 2000. Early Fixation of an Optimal Genetic Code. Molecular Biology and Evolution 17 (4), 511–518. https://doi.org/10.1093/oxfordjournals.molbev.a026331.
Friedman, S.M., Weinstein, I.B., 1964. Lack of fidelity in the translation of polyribonucleotides. Proceedings of the National Academy of Sciences of the United States of America 52 (4), 988–996. https://doi.org/10.1073/pnas.52.4.988.
Fukuda, Y., Tomita, M., Washio, T., 1999. Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumonia*. Nucleic Acids Research 27 (8), 1847–1853. https://doi.org/10.1093/nar/27.8.1847.
Hücker, S.M., Vanderhaeghen, S., Abellan-Schneyder, I., Scherer, S., Neuhaus, K., 2018a. The novel anaerobiosis-responsive overlapping gene *ano* is overlapping antisense to the annotated gene ECs2385 of *Escherichia coli* O157:H7 Sakai. Frontiers in Microbiology. https://doi.org/10.3389/fmicb.2018.00931.
Hücker, S.M., Vanderhaeghen, S., Abellan-Schneyder, I., Wecko, R., Simon, S., Scherer, S., Neuhaus, K., 2018b. A novel short L-arginine responsive protein-coding gene (laoB) antiparallel overlapping to a CadC-like transcriptional regulator in *Escherichia coli* O157:H7 sakai originated by overprinting. BMC Evolutionary Biology 18 (21). https://doi.org/10.1186/s12862-018-1134-0.
Haig, D., Hurst, L.D., 1991. A quantitative measure of error minimization in the genetic code. Journal of Molecular Evolution 33 (5), 412–417. https://doi.org/10.1007/BF02103132.
Higgs, P.G., Pudritz, R.E., 2009. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. Astrobiology 9 (5), 483–490. https://doi.org/10.1089/ast.2008.0280.
Itzkovitz, S., Alon, U., 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Research 17 (4), 405–412. https://doi.org/10.1101/gr.5987307.
Iwabe, N., Miyata, T., 2001. Overlapping genes in parasitic protist Giardia lamblia. Gene 280 (1–2), 163–167. https://doi.org/10.1016/S0378-1119(01)00767-3.
Kim, W., Silby, M.W., Purvine, S.O., Nicoll, J.S., Hixson, K.K., Monroe, M., Nicora, C.D., Lipton, M.S., Levy, S.B., 2009. Proteomic detection of non-annotated protein-coding genes in *Pseudomonas fluorescens* Pf0-1. PLOS ONE. https://doi.org/10.1371/journal.pone.0008455.
Knight, R.D., Freeland, S.J., Landweber, L.F., 2001. Rewiring the keyboard: evolvability of the genetic code. Nature Reviews Genetics 2, 49–58. https://doi.org/10.1038/35047500.
Konecny, J., Eckert, M., Schöniger, M., Hofacker, L.G., 1993. Neutral adaptation of the genetic code to double-strand coding. Journal of Molecular Evolution 36, 407–416. https://doi.org/10.1007/BF02406718.
Kumar, S., 1996. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. Genetics 143 (1), 537–548.
Ádám Kun, Ádám Radványi, 2018. The evolution of the genetic code: Impasses and challenges. BioSystems 164, 217–225. https://doi.org/10.1016/j.biosystems.2017.10.006.
Lynch, M., Marinov, G.K., 2015. The bioenergetic costs of a gene. Proceedings of the National Academy of Sciences United States of America 112 (51), 15690–15695. https://doi.org/10.1073/pnas.1514974112.
Makalowska, I., Lin, C.-F., Makalowski, W., 2005. Overlapping genes in vertebrate genomes. Computational Biology and Chemistry 29 (1), 1–12. https://doi.org/10.1016/j.compbiolchem.12.2004.006.
Martinez-Rodriguez, L., Erdogan, O., Jimenez-Rodriguez, M., Gonzalez-Rivera, K., Williams, T., Li, L., Weinreb, V., Collier, M., Chandrasekaran, S.N., Ambroggio, X., Kuhlman, B., 2005. Functional class I and II amino acid-activating enzymes can be coded by opposite strands of the same gene. The Journal of Biological Chemistry 290 (32), 19710–19725. https://doi.org/10.1074/jbc.M115.642876.
Massey, S.E., Sequential, A., 2006. "2-1-3" model of genetic code evolution that explains codon constraints. Journal of Molecular Evolution 62 (6), 809–810. https://doi.org/10.1007/s00239-005-0222-0.
Massey, S.E., 2008. A neutral origin for error m inimization in the genetic code. Journal of Molecular Evolution 67 (5), 510–516. https://doi.org/10.1007/s00239-008-9167-4.
Massey, S.E., 2010. searching of code space for an error-minimized genetic code via codon capture leads to failure, or requires at least 20 improving codon reassignments via the ambiguous intermediate mechanism. Journal of Molecular Evolution 70 (1), 106–115. https://doi.org/10.1007/s00239-009-9313-7.
Mathew, D.C., Luthey-Schulten, Z., 2008. On the physical basis of the amino acid polar requirement. Journal of Molecular Evolution 66 (5), 519–528. https://doi.org/10.1007/s00239-008-9073-9.
Mir, K., Schober, S., 2014. Investigation of genetic code optimality for overlapping protein coding sequences. 8th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)s 152–156. https://doi.org/10.1109/ISTC.2014.6955104.
Moriyama, E.N., Powell, J.R., 1997. Synonymous substitution rates in Drosophila: Mitochondrial versus nuclear genes. Journal of Molecular Evolution 45 (4), 378–391. https://doi.org/10.1007/PL00006243.
Morton, B.R., 1995. Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. Proceedings of the National Academy of Sciences of the United States of America 92 (21), 9717–9721. https://doi.org/10.1073/pnas.92.21.9717.
Parker, J., 1989. Errors and alternatives in reading the universal genetic code. Microbiology Reviews 53 (3), 273–298.
Rodin, S.N., Ohno, S., 1995. Two types of aminoacyl-trna synthetases could be originally encoded by complementary strands of the same nucleic ACID. Origins of life and evolution of the biosphere 25 (6), 565–589. https://doi.org/10.1007/BF01582025.
Sieber, P., Platzer, M., Schuster, S., 2018. The definition of open reading frame revisited. Trends in Genetics 34 (3), 167–170. https://doi.org/10.1016/j.tig.2017.12.009.
Spencer, C.A., Gietz, R.D., Hodgetts, R.B., 1986. Overlapping transcription units in the dopa decarboxylase region of Drosophila. Nature 322, 279–281. https://doi.org/10.1038/322279a0.
Tripathi, S., Deem, M.W., 2018. The standard genetic code facilitates exploration of the space of functional nucleotide sequences. Journal of Molecular Evolution 86 (6), 325–339. https://doi.org/10.1007/s00239-018-9852-x.
Tunca, S., Barreiro, C., Coque, J.R., Martín, J.F., 2009. Two overlapping antiparallel genes encoding the iron regulator DmdR1 and the Adm proteins control sidephore and antibiotic biosynthesis in Streptomycescoelicolor A3(2). The FEBS Journal 276 (17), 4814–4827. https://doi.org/10.1111/j.1742-4658.2009.2.x0718.
Vanderhaeghen, S., Zehentner, B., Scherer, S., Neuhaus, K., Ardern, Z., 2018. The novel EHEC gene *asa* overlaps the TEGT transporter gene in antisense and is regulated by NaCl and growth phase. Scientific Reports 8 (17875). https://doi.org/10.1038/s41598-018-35756-y.
Williams, T., Fried, M., 1986. A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3' ends. Nature 322, 275–279. https://doi.org/10.1038/322275a0.
Woese, C.R., Dugre, D.H., Saxinger, W.C., Dugre, S.A., 1966. The molecular basis for the genetic code. Proceedings of the National Academy of Sciences of the United States of America 55 (4), 966–974. https://doi.org/10.1073/pnas.55.4.966.
Woese, C.R., 1965. On the evolution of the genetic code. Proceedings of the National Academy of Sciences of the United States of America 54 (6), 1546–1552. https://doi.org/10.1073/pnas.54.6.1546.
Yarus, M., Widmann, J.J., Knight, R., 2009. RNA-amino acid binding: A stereochemical era for the genetic code. Journal of Molecular Evolution 69, 406–429. https://doi.org/10.1007/s00239-009-9270-1.
Zull, J.E., Smith, S.K., 1990. Is genetic code redundancy related to retention of structural information in both DNA strands? Trends in Biochemical Sciences 15 (7), 257–261. https://doi.org/10.1016/0968-0004(90)90048-G.