BMC Genomics

RESEARCH ARTICLE

Open Access

Biological factors in the synthetic construction of overlapping genes

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

Abstract

Background: Overlapping genes (OLGs) with long protein-coding overlapping sequences are disallowed by standard genome annotation programs, outside of viruses. Recently however they have been discovered in Archaea, diverse Bacteria, and Mammals. The biological factors underlying life's ability to create overlapping genes require more study, and may have important applications in understanding evolution and in biotechnology. A previous study claimed that protein domains from viruses were much better suited to forming overlaps than those from other cellular organisms - in this study we assessed this claim, in order to discover what might underlie taxonomic differences in the creation of gene overlaps.

Results: After overlapping arbitrary Pfam domain pairs and evaluating them with Hidden Markov Models we find OLG construction to be much less constrained than expected. For instance, close to 10% of the constructed sequences cannot be distinguished from typical sequences in their protein family. Most are also indistinguishable from natural protein sequences regarding identity and secondary structure. Surprisingly, contrary to a previous study, virus domains were much less suitable for designing OLGs than bacterial or eukaryotic domains were. In general, the amount of amino acid change required to force a domain to overlap is approximately equal to the variation observed within a typical domain family. The resulting high similarity between natural sequences and those altered so as to overlap is mostly due to the combination of high redundancy in the genetic code and the evolutionary exchangeability of many amino acids.

Conclusions: Synthetic overlapping genes which closely resemble natural gene sequences, as measured by HMM profiles, are remarkably easy to construct, and most arbitrary domain pairs can be altered so as to overlap while retaining high similarity to the original sequences. Future work however will need to assess important factors not considered such as intragenic interactions which affect protein folding. While the analysis here is not sufficient to guarantee functional folding proteins, further analysis of constructed OLGs will improve our understanding of the origin of these remarkable genetic elements across life and opens up exciting possibilities for synthetic biology.

Background

The triplet nature of the standard genetic code and double-stranded configuration of DNA together entail that six amino acid sequences are conceptually encoded within any nucleotide sequence, in different reading frames. Redundancy in the code mapping, whereby 64 nucleotide triplets encode just 20 amino acids, allows flexibility in what nucleotides are used for an amino acid sequence and consequently some freedom in the alternative frame sequences. The phenomenon of alternative frame coding has long been known to be utilised in viruses, with more than one amino acid sequence actually expressed from some loci [1, 2]. Unexpectedly however there is increasing evidence for overlapping coding in single-celled and multicellular organisms, for instance

*Correspondence: zachary.ardern@sanger.ac.uk
[1] Chair of Microbial Ecology, Department of Molecular Life Sciences, Technical University of Munich, Freising, Germany
Full list of author information is available at the end of the article

recent discoveries of fully embedded protein-coding genes encoded in alternate reading frames of known genes in Archaea [3], Bacteria [4–6] and even in mammals, including humans [7, 8]. Despite increasing evidence for their abundance, they are still generally not considered a significant phenomenon outside of viruses, due perhaps to perceived difficulties in their evolution for some or all reading frames [9, 10]. The idea that they are widespread has long been theorized however [10, 11]. Most gene prediction algorithms still exclude non-trivially overlapping genes [12], especially outside of bacteriophages and other viruses. The NCBI rules for annotation of prokaryotic genes do not allow genes completely embedded in another gene in a different frame without individual justification [13]. Even in viruses, relatively few fully embedded overlapping genes have been annotated, although more have recently been discovered, including in the pandemic viruses HIV and SARS-CoV-2 [2, 14, 15]. This ubiquity of overlapping coding has potential applications in biotechnology but is not well understood, and the contributions of fundamental biological factors such as the genetic code require further research.

Why do overlapping genes exist? While a mutation in a stop codon can easily create a short, trivial overlap in downstream neighbouring genes as a chance event, longer, non-trivial overlaps are only expected to be maintained over the long term if the overlapping region encodes a functional part of both proteins. There are a few hypothesised advantages to gene overlap, and the evidence for functional antisense overlaps and other alternative translation initiation sites in prokaryotes has been discussed in recent reviews [16, 17]. A selective advantage for long gene overlaps via the benefits of genome reduction has been proposed in viral genomes [18], but more recent evidence suggests that the genome reduction hypothesis has limited explanatory power [19]. Effects on gene regulation [20], in contrast, could be relevant across taxonomic domains. For instance, two genes in a same-strand overlapping gene (OLG) pair could perhaps be efficiently co-expressed if encoded within the same mRNA. Genes within an antisense overlapping pair could also influence each other, in a way similar to what has recently been termed a "noncontiguous operon" where non-overlapping genes encoded in antisense to each other were observed to be co-expressed as an operon [21]. A less direct proposed benefit of overlapping genes concerns gene origin - ORFs (open reading frames) overlapping existing genes may become translated and give rise to new genes [22, 23]. One aspect of this is that sometimes protein structure of new genes arising from overlapping ORFs may be partially templated from the existing 'mother gene'. For genes encoded directly in antisense ("-1 frame") there is a tendency for the creation of proteins with a complementary polarity structure to the gene on the antisense strand [24–26]; while in the case of same strand, or sense-sense overlaps a similar hydrophobicity profile between unshifted and shifted frames has been observed [27, 28]. In general, overlapping open reading frames may play an important role in the origin of *de novo* genes, exploring new territory in the total space of sequences and functions [23, 27, 29–32]. While most currently extant OLGs in viruses are not taxonomically conserved and therefore appear to be evolutionarily young [33], one claimed example of an ancient OLG pair in cellular organisms is comprised of the two classes of aminoacyl-tRNA synthetases which can be encoded in an overlapping manner [34–38]. An ancient pairing between NAD-glutamate dehydrogenase and a heat shock protein 70 has also been proposed [39, 40], but has been a topic of controversy [41].

A previous study by Opuu, Silvert, and Simonson [42] quantified the difficulty of constructing OLGs by picking random pairs of protein domains and rewriting them so as to overlap, with an algorithm minimizing the amino acid changes in each domain. This was a new approach, as prior studies had tried to create overlaps without changing the amino acid sequence of the two genes, which resulted in either a very limited overlap length [43] or could only be done for very specific genes [44]. They found that, remarkably, 16% of 125,250 arbitrary protein domain pairs were able to successfully overlap in at least one of the 3 reading frames they investigated, and one of two positions tested. Virus domains were much more likely to create putatively functional overlaps than were domains from prokaryotes or eukaryotes, as determined by BLAST searches of the SWISS-PROT database. This result suggests that creating overlaps is not as difficult as might be expected, implying that an abnormally high threshold of evidence as compared to other gene types should not be required in order to verify their existence. In our study the algorithm provided in [42] is improved in the evaluation of the constructed sequences as the previous analysis had some weaknesses resulting in incorrect claims. Determining whether an artificial sequence has a specific function based on its amino acid sequence only is a very hard problem and not possible today. Remarkable progress is being made in predicting the protein structure from amino acid sequence [45–47], but protein structure does not determine function as essential binding sites can be rendered useless if the amino acid is changed even when the overall protein structure remains the same. Ultimately only laboratory experiments can definitively determine the function of a given amino acid sequence. In order to aid the design of expensive experimental setups however, it can at least be determined bioinformatically how similar an artificial sequence is to

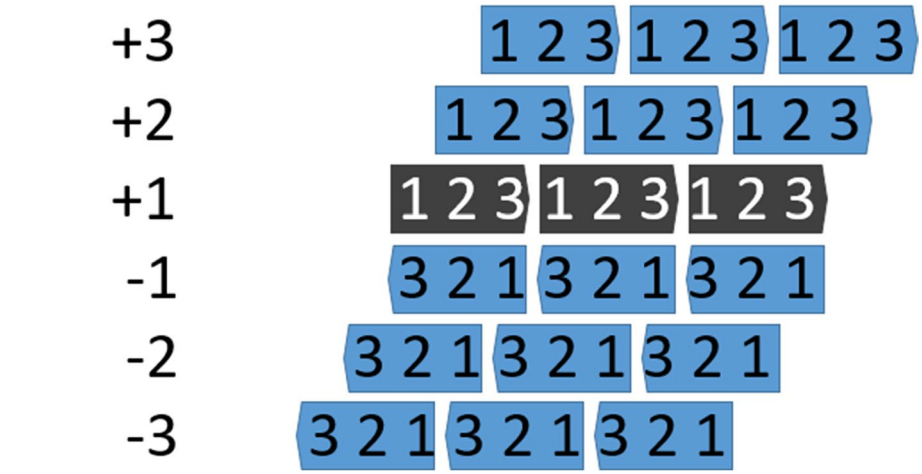

Fig. 1 Illustration of the alternative reading frames. The '+1' frame is the standard or reference reading frame and '+2'/'+3' the sense overlaps, while frames '-1' to '-3' are on the anti-sense strand. Figure from Wichmann and Ardern (2019) [51]

sequences with known functions. In this study the artificially designed sequences are compared to their original sequences in terms of amino acid identity, amino acid similarity, Hidden Markov Model profile and secondary structure in order to determine the impact of OLG construction and which sequences are potentially functional. We note that another recent study [48], independently made similar advances to this study in the use of a HMM algorithm for assessing the quality of designed overlaps, but also added a more sophisticated step which takes into account intra-protein interactions. While the study was multifaceted, Blazejewski et al. did not make comparisons across taxonomic groups or genetic codes, or rigorously assess success rates as we have done here.

In our study, we fully overlap arbitrary pairs of natural domains from the Pfam database, at random positions, re-implementing the algorithm of Opuu et al. [42] and expanding it to all alternative reading frames. A Hidden Markov Model approach rather than BLAST is used to assess the quality of designed OLGs. Also, while an estimate for an upper limit on how many domains can be successfully overlapped in at least one reading frame and position was previously made [42], here the average success rate for OLG construction is determined instead. This is more relevant for understanding constraints on the formation rate of naturally occurring OLGs and in assessing the probability of successful synthetic creation of OLGs. Our results in one sense give an upper estimate of the ease of creating overlaps as the difficulty of obtaining an overlapping gene pair naturally is not directly addressed here. On the other hand, overlapping functional domains directly is a "worst case scenario" as there is some evidence that the critical functional domains of one protein in an OLG pair tend to overlap less constrained regions or residues of the other protein [49, 50], and this segregation is also intuitively plausible. However, a major limitation of the approach used here (in common with the previous study on which the analysis is based - [42]) is that HMM profile methods do not take into account crucial interactions within a protein. As such, whenever "success" is used in this study with regards to a constructed overlap it cannot be taken to entail biological functionality, but rather just entails having met an important first step in constructing protein-like sequences.

In order to estimate the difficulty of achieving overprinting naturally, the minimal number of nucleotide changes needed to create the OLG sequence is determined. Whether functional domains do in fact overlap in nature, however, deserves further attention. Because we expand the analysis of Opuu et al. [42] from the reading frames '+2',-1' and '-3' to all reading frames (see Fig. 1 for reading frame definitions), we are able to relate the observed differences between reading frames to the structure of the standard genetic code. Potential optimisation of the standard code for OLGs can be studied by constructing OLGs using randomly generated genetic codes. Using the improved evaluation of the designed OLGs we study the impact on properties of constructed domains in terms of amino acid identity, similarity, and protein secondary structure. We also investigate the evolutionary accessibility of the constructed domains in

terms of total sequence change required, and what influence filtering to proteins restricted to particular taxonomic groups (domains) has on the results.

Results

Previous dataset-database biases and length-dependence

In the previous study [42], constructed sequences were evaluated with a BLAST search against the SWISS-PROT database. If both overlapping sequences had a match to the best hit with at most an e-value of 10^(-10) and a match length of 85%, the overlap was considered successful. However, the initial sequences were picked from the Pfam seed database and it can be shown that most of the chosen sequences are not well represented in the SWISS-PROT database (see left panel in Fig. 2), with the exception of virus genes. In a search against the SWISS-PROT database, identities of over 80% were only found for 15% of the non virus genes, while 70% of the virus genes could be found in this category. A curated set in which all sequences have a 100% match in the SWISS-PROT database but otherwise the same properties has a remarkable 95% success rate for overlaps and the virus vs. non-virus difference vanishes (see right panel in Fig. 2). The advantage reported for virus genes is thus fully explained by dataset-database biases. In either case however, the extremely high overall success rate obtained should be investigated. Either creating overlaps is indeed unexpectedly easy or the evaluation of functionality used in [42] is not conservative enough. It can be shown that both factors appear to contribute to the surprising result.

Another difficulty with the previously published approach is that expectation values, e-values, are expected to be length-dependent. For instance, when introducing the minimal number of changes required for two random sequences to fully overlap each other, a similar percentage of each sequence is expected to change. As a longer sequence with the same similarity has a lower probability of being found by chance in a database of a given size, in such a case the e-values of the constructed sequences would be strongly dependent on the length of the input sequences. Such a length-dependence can be found in the BLAST evaluation (Supplementary Fig. S1). A fixed e-value cutoff cannot adequately evaluate sequences in such a situation as the cutoff value fully determines the result and is chosen arbitrarily. The sequences used in Opuu et al. [42] have a length of 70-100 amino acids, and the observed high success rate for the curated set can be explained by a combination of the sequence length and the choice of the cutoff value.

Advantages of using Hidden Markov Models over BLAST

In order to find a reasonable alternative to the fixed e-value cutoff for construction success, a relative threshold is calculated through comparison to a set of

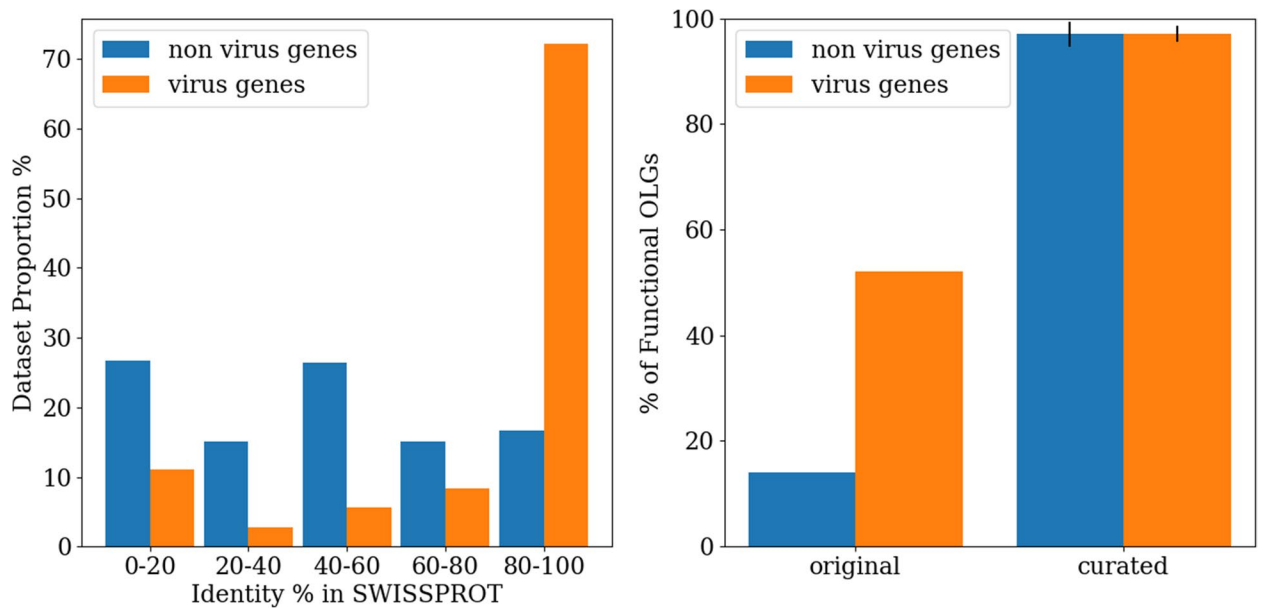

Fig. 2 *Left*: Proportions of the dataset used in Opuu et al. with different match identities in SWISS-PROT - virus genes from this dataset have a higher average identity to a SWISS-PROT entry than non-virus genes. *Right*: Percentage of functional OLGs for the original dataset from Opuu et al. and the average of 10 curated datasets grouped into virus and non virus genes. In curated datasets all original sequences have an exact match in SWISS-PROT. Each curated dataset has 100 sequences with 70-100 amino acids. The virus versus non-virus difference observed in the previous study's dataset vanishes for the curated datasets

homologues of the original sequence, obtained from the Pfam database [52], which is divided into curated sets of "seed" sequences which are the basis of each domain family, and the full database which is clustered according to similarity to the seed sequences. In our approach, a constructed sequence is subsequently judged successful if its score against the HMM profile is higher than a certain percentile of the 'full' sequences in the appropriate family, thereby creating a threshold value which is individual for each protein family and mostly independent of the length of the domains (see Methods). When judging the success of a constructed sequence, two particular percentile values are referred to in relation to the natural sequences in the Pfam domain family. Firstly, the 50th percentile (median), is used to mark the score of a typical sequence in the protein family. Sequences meeting this threshold can not be distinguished from the naturally occurring protein domains with HMM profiles and they will be categorized as typical proteins. Secondly, since all of the homologues used are naturally occurring sequences, scoring at least as highly as any one of these sequences renders a sequence what we term "biologically relevant". In order to avoid extreme outliers which may be misclassified however, the score of the 5th percentile of a domain family is conservatively used as the biologically relevant threshold for that family.

The construction algorithm previously used takes the conservation of each position in the protein domains into account in order to find an optimal OLG pair. Introducing these positional weights results in a lower success rate in the subsequent BLAST evaluation for OLG quality or success, since BLAST compares whole sequences pairwise and does not take the conservation of different positions into account. Using HMMs for assessment on the other hand results in an increased success rate when introducing positional weights. Here we also optimise the strength of these positional weights (see methods).

Retaining amino acid sequence similarity

Retention of amino acid sequence identity is another - albeit minimal - indicator of functionality. It has been argued that a 34% amino acid identity between naturally occurring sequences ensures that both sequences have the same structure [53]. Comparing the altered part due to OLG construction with the original sequence, in 96.5% of cases both OLG sequences share at least 34% of amino acids with their original sequence. In some OLG pairs both sequences have an amino acid identity of up to 60% compared to their original sequence. In the arguably biologically more relevant property of amino acid 'similarity', the worst-scoring of the two OLGs can be even up to 80% similar to its respective original sequence (cf. left panel of Fig. 3).

One quantitative impact of OLG design can be measured via the average amino acid identity and similarity between the two OLG sequences. The average amino acid identity is 60% in most cases (right panel of Fig. 3);

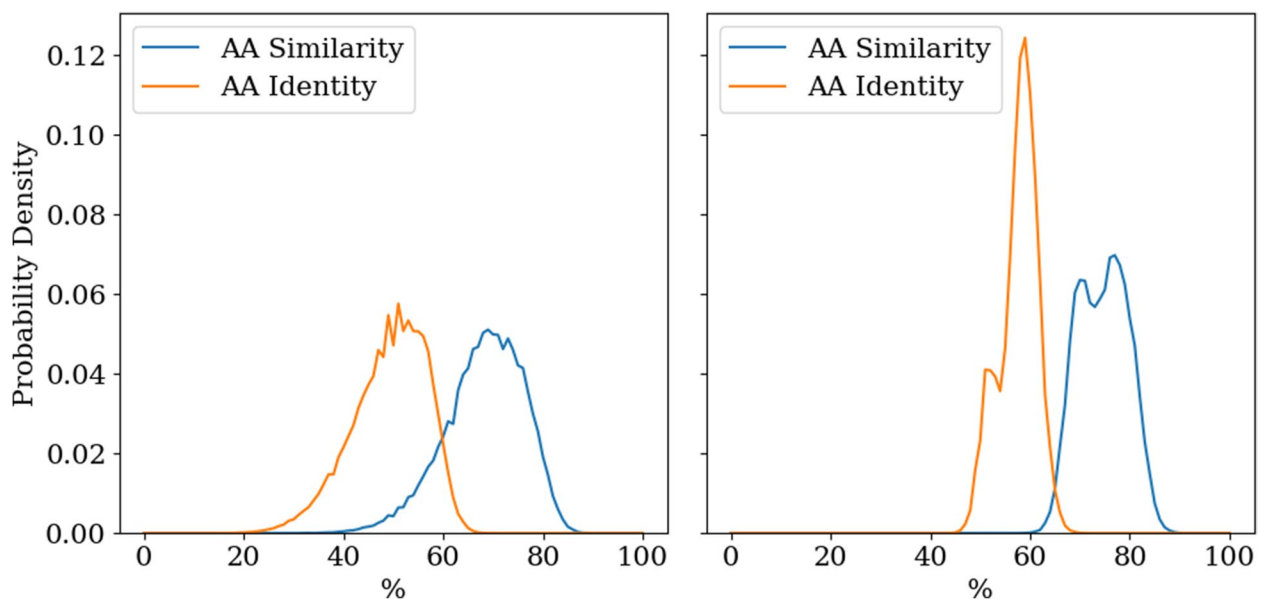

Fig. 3 Probability density for different amino acid identities and similarities in constructed OLG pairs. The data is calculated from 505,000 OLG pairs. *Left*: The sequence with the lower identity is representative of the pair. The black line indicates the 34% amino acid identity threshold. *Right*: The mean similarity of both OLG sequences represents the pair

in almost all OLG pairs one sequence is above and one is below 60% amino acid identity. The average amino acid similarity is 75% in most cases (right panel of Fig. 3) - again in almost all cases one of the two OLG sequences is above and one below 75% identity. The double peak structure of both panels in Fig. 3 can be explained by differences for OLG pairs in different relative reading frames, which are pooled here (cf. Supplementary Fig. S2). It follows that for an average designed OLG pair, it can be estimated that in 20% of all overlap positions the amino acids of both sequences are maintained, in 30% one sequence maintains its amino acid while the amino acid in the other sequence is changed to a similar one and in 50% one sequence maintains its amino acid and the other sequence cannot maintain a similar amino acid. Precisely how well the two sequences can be maintained after designed overlap is determined by the standard genetic code, the two specific sequences, the overlap position, their amino acid composition and the amino acid order. While the standard genetic code is a constant factor across all overlaps, all other factors are specific in each case and create the observed variability in the results.

The impact of OLG construction on secondary structure is the last factor studied here. Comparing the predicted secondary structure of the OLG sequence with that of the natural non-overlapped sequence, a secondary structure similarity score is determined. Secondary structure is predicted using Porter 5 [54] with the "--fast" flag. This program can distinguish between the eight different secondary structure motifs of the dictionary of protein secondary structure (DPSS) [55–57], which are 3_10-, alpha-, and phi- helices, hydrogen bonded turns, beta sheets, beta bridges, bends and coils. Determining the same secondary structure similarity for all sequences in the seed group of the Pfam database yields a control group. This way the typical deviations between domains with the same function can be determined. Comparing probability densities for different secondary structure identities in both groups it can be seen that the constructed OLG sequences barely deviate from the seed sequences (cf. Figure 4). In conclusion, in regards to secondary structure the change inflicted on a sequence to create OLGs is no more than the differences within naturally occurring protein domain families.

It is noteworthy that only amino acid identity and similarity have a strong correlation (r=0.82) so combined with the other parameters, namely the relative HMM score (see methods) and the secondary structure identity, there is a set of three more or less independent properties for evaluating constructed OLGs, and probably for protein homologues in general. The relative HMM score is the HMM score of the OLG sequence divided by the HMM score of a sequence at any threshold percentile as discussed above in the section on the advantages of HMMs. Between each pair of parameters the Pearson's correlation is below 0.2, with the exception of the correlation between secondary structure identity and HMM score which has r=0.37 or r=0.39 for thresholds of 95% or 100% respectively.

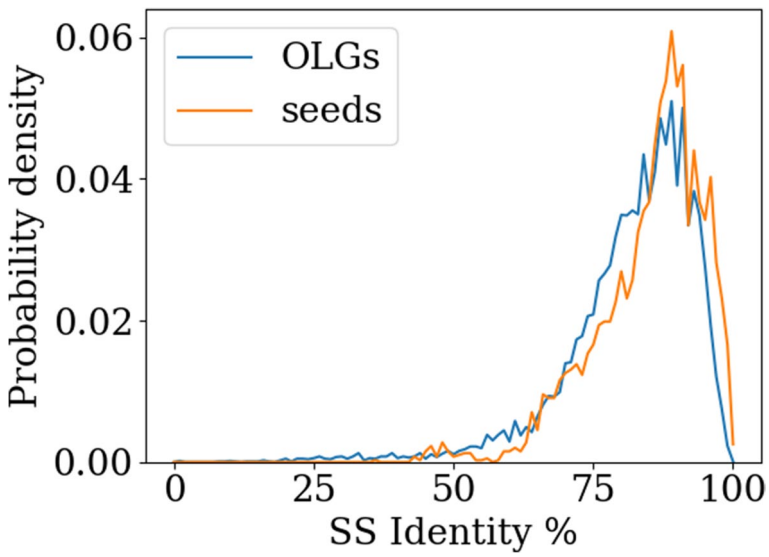

Fig. 4 Probability densities for different secondary structure identities for OLGs and seed sequences calculated from a dataset of 50 sequences consisting of at least 70 amino acids. OLGs are as similar to their original sequences in secondary structure as observed for comparisons of seed sequences of naturally occurring protein domains to the sequence best representing the respective domain family

Construction success rates and reading frame

The proportion of potential overlap positions where successful overlap is achieved differs across genes. The distribution of the percentage of successful positions in each OLG pair was calculated from up to 50 different positions across all possible pairings of 150 domains (see Fig. 5). 50.3% of all OLG pairs form biologically relevant sequences (with the worst-scoring sequence of the pair scoring at least as highly as the lower 5% of the original sequence's domain family) at all positions in every reading frame while only 2.5% cannot form a biologically relevant sequence at any position (see Fig. 5). 1.9% of the pairs even form typical proteins, as determined by the 50th percentile threshold, at every position in any reading frame (see right panel in Fig. 5). This result is strongly dependent on the threshold percentile chosen, but due to the wide range of possible results it can still be concluded that the chance of success of a constructed OLG pair depends strongly on the particular genes used, as might be expected.

The five alternative reading frames differ strongly in the combinatorial constraints imposed by the reference gene (mother gene) via the standard genetic code [9], e.g. the sequence N|GCN|, with N being any nucleotide, always translates to alanine in both the +1 and the -2 frames. We investigate whether this difference in constraint transfers to the success rate for designing OLGs. For OLGs resembling typical proteins of their respective families, the success rates for OLG construction by our analysis varies from 14.9% in the '-3' frame to 3.0% in the '-2' frame with an average value of 9.6% across all reading frames (see Fig. 6). Calculating the e-value just as in the earlier study [42] as a reference, the constructed OLGs have a median e-value of 10^-(28) to 10^(-37), decreasing with stricter homolog family percentile. The result is strongly threshold (homolog family percentile) dependent as 94.5% of the constructed sequences score at least as highly as the worst sequence in the full group, while only 0.02% score better than 95% of the full group. Considering combinatorial restrictions of different reading frames, the rankings of frames by success rate are exactly as expected, insofar as the success rate of each reading frame is inversely proportional to the extent of combinatorial restrictions calculated in Lèbre and Gascuel [9] (see Fig. 6): the '-2' frame is the least successful reading frame and has the highest restrictions, followed by the '-1' frame, which is the second most restricted frame. Next are reading frames '+2' and '+3', which have exactly the same restrictions and surprisingly almost the same success rates, not only in their average value but also in every single dataset (data not shown), despite expected stochastic fluctuations due to some genes simply fitting better to each other. Last is the '-3' frame, which has no combinatorial restrictions and the highest success rate. Plotting the different success rates in the different reading frames as a function of the calculated number of combinatorial

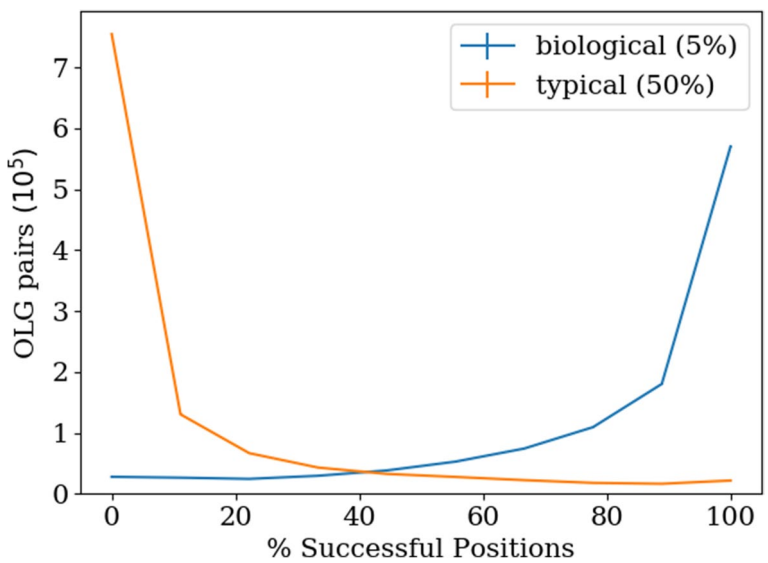

Fig. 5 Frequency of successful overlap positions in OLGs. 150 randomly chosen domains with a minimum length of 70 amino acids are used as a basis, resulting in 11,325 OLG pairs. In each OLG pair 30 sets of up to 50 random positions were tested against the Pfam group HMMs using the 'biologically relevant' threshold (5th percentile) and the 'typical sequence' threshold (50th percentile) for a successful overlap. While 50.3% of the pairs can be overlapped at any position and 2.5% in no position using the biological threshold only 1.9% can be overlapped at any position and 66.7% in no position using the threshold of typical sequences. The sequence threshold strongly influences the result

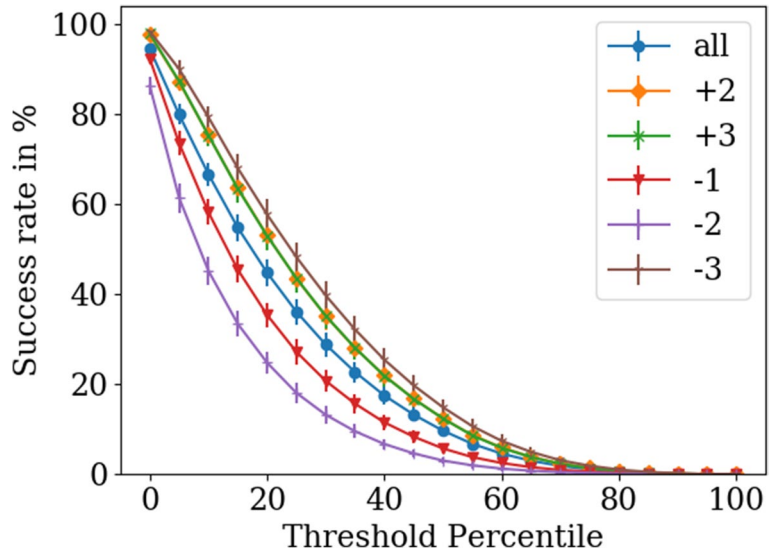

Fig. 6 Success rates for OLG design in different reading frames as a function of threshold percentile. Each value is an average from 20 different datasets of 150 sequences with at least 70 amino acids and the error bars are equal to the standard deviation. The threshold chosen within the Pfam group has a very strong influence on success rates. The ordering of the reading frames by success rates, namely '-3', '+2'/'+3', '-1' and '-2', matches the ordering by combinatorial restrictions in the standard genetic code, beginning with the least restricted frame

constraints [9], results in a linear relation for the lowest possible threshold, namely that all sequences which are at least as good as the worst in the comparison group are judged successful. As the threshold is increased the linear relation is gradually lost (see supplementary Fig. S3). For higher thresholds most of the sequences are below the threshold and very little data is left, which might lead to the observed behaviour. In summary, the structure of the standard genetic code appears to strongly influence the construction of OLGs. Whether the observed relationship between predicted constraints in different frames and the difficulty of constructing OLGs is borne out by the proportion of natural OLGs found across frames deserves attention across diverse taxa.

Taxonomic differences

Besides the four basic taxonomic groups (three domains of cellular life: archaea, bacteria, eukaryotes, plus viruses) also ancient genes can be studied by picking only families which have at least one sequence in all four taxonomic groups since it is expect that these families have already been present in LUCA or another ancient ancestor (although this high level categorisation is not perfect due to widespread horizontal gene transfer). Surprisingly, bacterial and eukaryotic genes are generally significantly better suited to OLG construction than virus and archaeal genes with only minimal dependence on the threshold percentile, cf. Fig. S4. The largest dependence on the threshold percentile is found for the "Found in All" genes, of which only a total of 50 sequences can be found in the Pfam database, so higher stochastic fluctuations are to be expected. Using the 'biologically relevant' threshold, the biggest difference is between eukaryotic and archaeal genes which have a 20% difference in their success rate (see left panel of Fig. 7). For OLGs which are typical proteins of their respective family, eukaryotic genes are almost twice as likely to be successful as virus genes (see right panel of Fig. 7).

Eukaryotes and "Found in All" genes are typically the easiest to overlap, which is somewhat unexpected as domain families restricted to eukaryotes might be expected to typically be younger, and so to appear less 'flexible' due to having sampled less of the functional sequence space through mutations. More understandable however is that due to being closer to mutational saturation (if more ancient on average) and therefore having explored a larger proportion of functional sequence space, "Found in All" genes might appear more 'flexible', resulting in lower weights and thresholds.

Evolutionary distance of constructed OLGs to biological sequences

In order to estimate the difficulty of naturally forming OLG sequences, the minimum number of nucleotide changes needed in order to reach the OLG sequence from any of the two original sequences is determined (see Fig. 8). By only taking OLGs in which both sequences are above a certain HMM threshold, extreme outliers

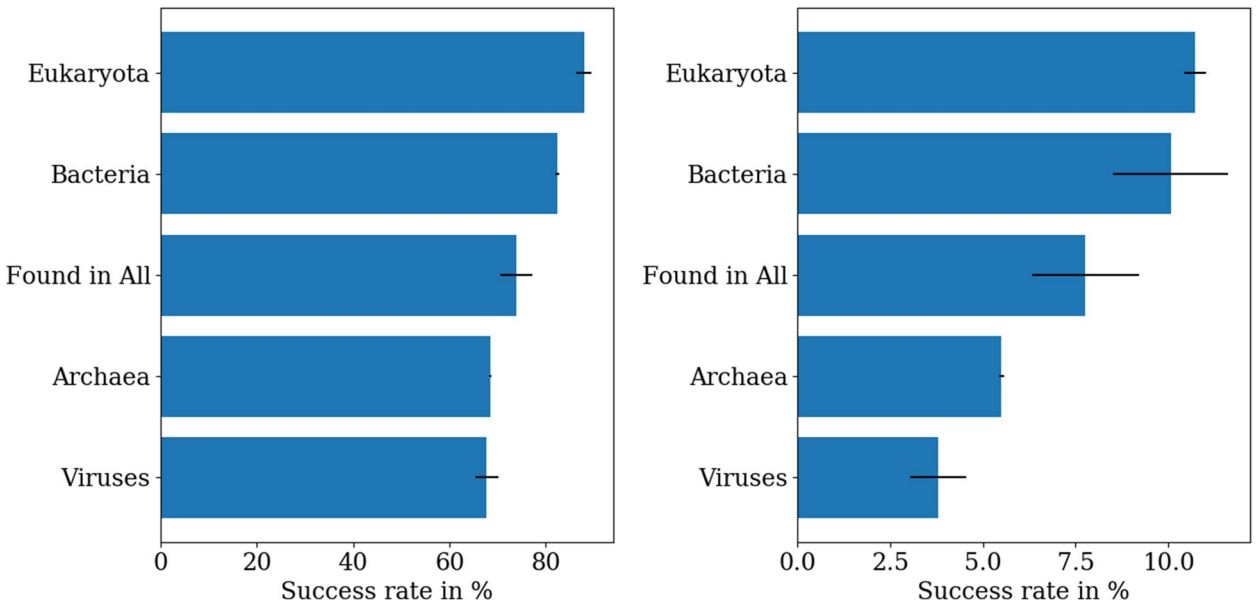

Fig. 7 Average success rates in different taxonomic groups calculated from 20 datasets with 150 domains consisting of at least 70 amino acids each; the black lines indicate the standard deviation. The ordering of 'biologically relevant' genes (threshold percentile 5%, *left*) is mostly identical to 'typical' genes (threshold percentile 50%, *right*). Sequences from Bacteria and Eukaryota sequences outperform Viruses and Archaea

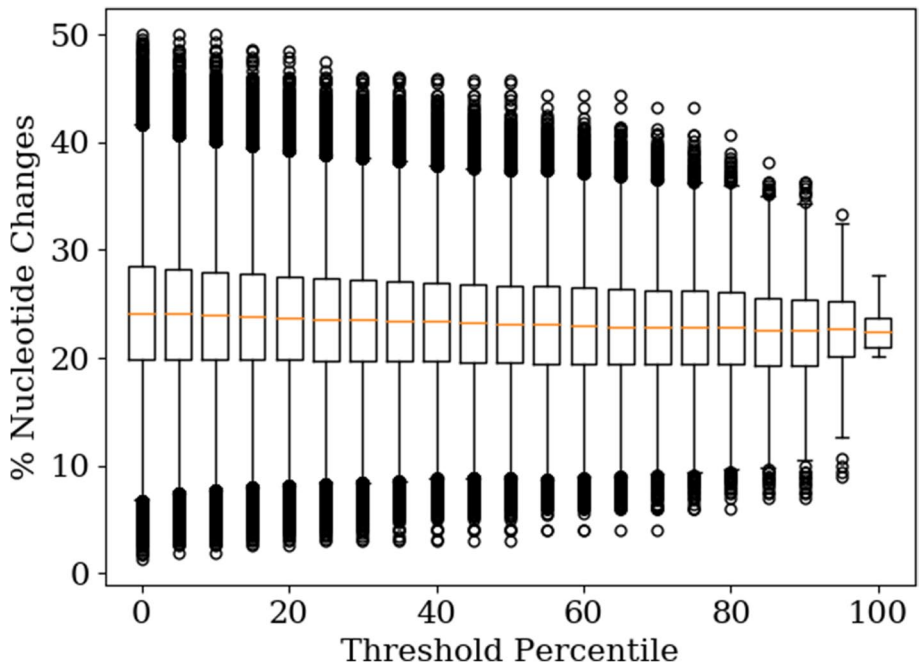

Fig. 8 Percentage nucleotide change of OLGs as a function of HMM threshold %. The minimal nucleotide distance of each of 1,010,000 OLG sequences (two per pair), with a minimal length of 210 nucleotides, to their respective original sequence is determined

are gradually removed with increasing threshold but the rest of the distribution stays the same. This indicates that this property is independent of the threshold value, just as for the amino acid identity and similarity, as fewer and fewer designed OLGs pass a higher threshold which makes extreme outliers less likely to occur. On average a designed OLG sequence has a 25% difference in nucleotides to its original, with half of constructed sequences in

the range of 20-30% change. Most interesting are outliers on the lower end of the distribution as they indicate whether OLGs exist that are potentially reachable by naturally occurring mutations. The lowest nucleotide difference observed is 1.8%, which was for an OLG pair that scores better than 25% of the domains in the comparison group. 0.6% of OLGs required less than 10% nucleotide change, i.e. 5843 sequences of the 955,846 sequences created in this dataset that scored at least as highly as the worst sequence in the comparison group. This suggests intuitively that creating overlaps of the sort constructed here could be possible naturally through accumulation of random mutations. The population genetics of such a hypothetical process is a potential topic for further study, as is an experimental evaluation of functionality.

Influence of the genetic code and code optimality

By comparing OLG sequences constructed with the standard genetic code (SGC) to sequences constructed with artificial codes the level of optimality of the SGC can be inferred. Since such an approach depends strongly on the codeset used [51], four different versions with increasing restrictions will be tested. There are two factors defining a genetic code, namely its amino acid composition and the arrangement of amino acids on the 64 codons. The first code set is the random code set and does not constrict any of the two factors. Each code can have any of the 20 amino acids used in the SGC at any codon. The second set only restricts the composition of its codes and is called the degeneracy code set. All codes in this set contain the same amount of codons for each amino acid as in the SGC and thus conserving its amino acid composition. The third set is the blocks code set whose codes have a very similar structure to the SGC and while it also restricts the composition of the codes to some degree it mostly determines their arrangement. This code set is created by assigning all codons of the SGC that code for the same amino acid into blocks and shuffling the amino acids assigned to each block and thus conserves the degeneracy structure of the SGC on the third nucleotide. Lastly a code set that maintains the mutational robustness of the SGC as previously calculated [51] is tested. In short, the mutational robustness is the average change of amino acids due to point mutations and has been shown to be extremely optimal in the SGC relative to similar codes [58]. This set contains block codes like in the second set but only the codes whose mutational robustness is at least as high as the SGC are kept. Since these codes are fundamentally block codes they are partly restricted in their amino acid composition, but the arrangement of amino acids in these codes is even more restricted as point mutations from any codons should result in similar amino acids. This code set reflects the fact that different properties of the SGC have a different impact on the fitness or biological optimality of the SGC with the mutational robustness most likely being one of the most important features. Here this code set is called the mutational robustness blocks set (MR-blocks set) and it tests the optimality of constructing OLGs as an additional property of the SGC after taking into account the mutational robustness.

Comparing the degeneracy, the block and the MR-blocks code set to the random set, the influence of code composition and arrangement can be determined (see left panel of Fig. 9). The degeneracy code set reflecting the composition of the SGC has the codes with the highest average success rates indicating that the composition of the SGC is a major factor for this property, but the SGC itself has a very low success rate in comparison, indicating that the amino acid arrangement is an even stronger - in this case negative - factor as the SGC is worse than both the random codes and the degeneracy codes. The block structure of the SGC has a strong negative impact on successful OLG design and the SGC is a typical member of this set. Enforcing even more structure on the artificial codes in order to maintain the mutation robustness of the SGC further reduces the ability of the SGC to create successful OLGs.

Studying the optimalities of each of the four code sets for flexibility in OLG design, it is apparent that the more restricted the code set is, the more optimal the SGC is relative to the set (see right panel of Fig. 9). Especially in the MR-blocks code set only a few codes are better than the SGC, however no codeset or reading frame has fewer than 5% of codes doing better (see Supplementary Figures S5-S8), which has been a recommended threshold for inferring optimality [59]. This is an expected result even if the code has been optimised for OLGs as the success rate for constructing OLGs reflects merely the 'flexibility' of a code system, but OLG sequences also need to be conserved, which is an almost directly opposing property which also has not been found to be strongly optimal by itself [51]; it might indeed be expected that overall optimality involves a trade-off between the two.

If the SGC has been optimized in this way this could indicate a turning point at which a further increase in mutational robustness results in a smaller fitness increase compared to an increase in the flexibility to create OLGs - how to measure fitness for a genetic code is however not clear. While the code composition of the SGC is beneficial for both the ability to create successful OLGs and the mutational robustness, the code arrangement of the SGC is only beneficial for mutational error robustness (see Fig. 2 of Wichmann and Ardern [51]), indicating that, in an optimization framework, the mutational robustness is the more important property. Only in the

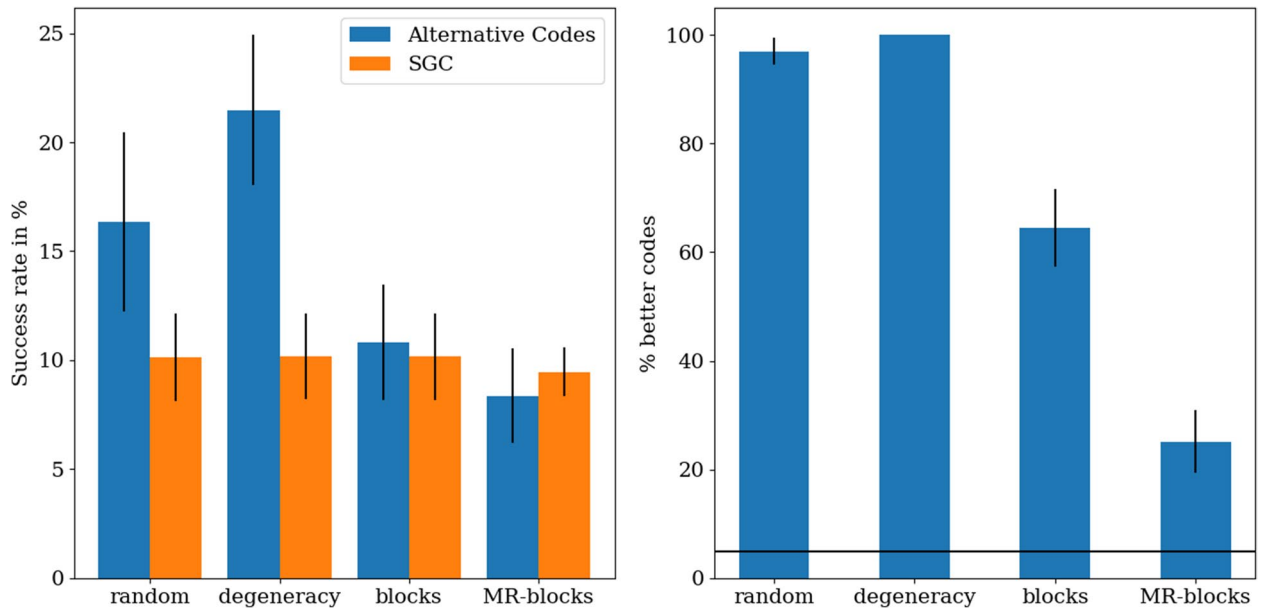

Fig. 9 OLG design success rates for different alternative code sets. The average is calculated from 20 sets of 100 alternative codes, except for the MR-blocks set with 10 sets of 500 codes. The error bars indicate the standard deviation. *Left*: The average success rates compared to the SGC. While the composition of the SGC is a positive factor, the arrangement of the SGC is a negative factor. *Right*: The optimality of different code sets. The black line indicates the 5% threshold. The more restricted the code set the more optimal the SGC appears indicating that the ability to successfully create OLGs can only be seen as optimized in the SGC if other properties are maintained

set of codes with the same mutational robustness does the optimality for OLG design become stronger, supporting the 'turning point' hypothesis.

Discussion

There are many different aspects of the synthetic construction of OLG pairs which could be studied for useful insight. Here factors including sequence length and the influence of sequence conservation are taken into account. The analysis shows that an evaluation with BLAST and a fixed e-value cutoff cannot accurately assess the potential functionality of the designed OLGs. While the combination of sequence length and an e-value cutoff completely determines the success rate of the constructed OLGs, adding in positional weights can only negatively influence the sequences constructed with this method. Both problems can be approached more fruitfully however by instead using HMM profiles to determine sequence similarity and then using these to define a threshold for successful OLGs derived from sequences in the same protein family. In addition, further optimization of the construction algorithm can be achieved by determining the optimal weight strength (influence of sequence conservation), which we found to be k=0.5. The HMM profiles and the thresholds are however both derived from the Pfam database [52], which makes these results strongly dependent on the database quality. For example, if in one taxonomic group sequences are very similar due to being mostly from the same species or genus, thresholds would appear to be higher and it would be harder for designed OLGs to pass these thresholds. More fundamentally, the underlying alignment processes used to create the database are imperfect, and more precise approaches such as including structural information during alignment [36] would improve accuracy - with the recent expansion of *in silico* structural determination [46], many opportunities for improved studies of protein families will be possible. Taking into account intra-protein interactions, as has been done in an impressive recent study of synthetic OLG construction [48], will also be important in future, as HMM methods treat each position in the alignment as independent.

We find that 94.5% of the constructed OLG sequences score at least as highly as the worst-scoring biological sequences in Pfam groups - i.e. the vast majority of constructed sequences would fit into the Pfam group of the natural sequence. Further, 9.6% of the sequences cannot be distinguished from naturally occurring domains in their respective protein family, in that they score better than 50% of family members. This gives the intriguing result that the typical variation inside protein families is of the same order of magnitude as the change needed in order to construct artificial OLGs by arbitrary pairing of protein domains. This result also holds true for

other bioinformatic factors like amino acid identity and secondary structure, since the constructed OLGs are typically very similar to naturally occurring domains in these properties. Studying artificial OLG design success from the perspective of an even more constraining biological parameter like tertiary structure would be an important next step, and has been implemented to some extent in a recent study [48]. Fully taking into account effects on tertiary structure is complex, as besides the amino acid sequence, codon usage can also impact protein structure [60], along with environmental factors such as the presence of chaperone proteins. Ultimately, proof of the functionality of artificial sequences cannot yet be realised bioinformatically, and experimental verification is required. To this end, ideally all known independent protein properties available from the sequence should be tested in order to create a gold standard for possibly functional sequences. From this study it is clear that sequence similarity (or identity), HMM-scores and secondary structure are relevant properties that all contribute something when judging similarity to natural sequences. Determining relative HMM scores for high thresholds could be used to prefilter sequences for secondary structure prediction as it is the computationally most intensive part of this analysis.

Considering that domain-domain overlaps are expected to be much harder than overlapping a domain with a less conserved region in another gene, it appears that de novo origin of genes from overlapping ORFs may be much less difficult than widely assumed. Some constructed OLG sequences varied only by 1.8% from their original sequence, and there will plausibly often be other natural sequences from the same domain family that are even closer to the OLG sequence. This result could be a starting point for estimating the difficulty of evolving OLGs from different starting sequences in natural systems, which is still relatively unexplored despite some early work [61]. The structure of the standard genetic code is crucial in explaining differences between reading frames and is a strong factor in the overall success rate of OLG construction. For example, OLGs can maintain an average 60% amino acid identity and an average 75% amino acid similarity, which is mostly due to the genetic code. The structure of the standard genetic code is defined by its composition, namely how many codons code for each amino acid, and its arrangement, namely which codons code for each amino acid. It is known that the composition alone cannot explain the strong optimality of the standard genetic code for mutational robustness as it stands out from between codes with the same composition as the standard genetic code [51, 62]. Considering that the arrangement of the standard genetic code creates such high mutational robustness values [58] it is remarkable that designing OLGs also works so well.

The analysis presented here depends primarily on the reliability of HMM profiles of Pfam groups as a guide to biological functionality in constructed sequences. Reliability for classifying biological protein sequences into ortholog families, the main use of these HMMs, may not correlate well with reliability in scoring artificially constructed sequences for functionality. In other words, these profiles no doubt fail to capture some important requirements for protein tertiary structure and/or functionality. Future research ought to test the best candidates experimentally, and if the best candidates from the methods developed here are not successful, additional factors should also be considered in comparing constructed sequences and their natural precursors. For instance, many protein characteristics can be assessed using servers or packages incorporating multiple bioinformatic tools such as PredictProtein, for various secondary structural elements [63], and many sequence properties, such as hydrophobicity profiles, can be computed using the VOLPES server [64], related methods to which have been applied to the related case of frame-shifted sequences compared to their mother genes [27]. Other properties required for functional protein sequences can be inferred from the evolutionary information contained in sequence alignments of protein families. For instance, it has been calculated based on a study of residue-residue co-evolution in ten well-characterized protein families that the proportion of all sequences which fold to the family's structure ranges from approx 10^{-24} to 10^{-126} [65]. These principles have been successfully used in the design of functional proteins [66], and could conceivably also be applied to OLG construction.

Factors facilitating the existence of OLGs may possibly help in predicting OLGs in sequenced genomes and should be explored further. For instance, a careful study of relatively 'flexible' sequence regions in taxonomically widespread genes may help find more overlapping genes. Most interestingly, bacterial and eukaryotic genes can be overlapped more easily than virus genes, contrary to previous findings [42]. These earlier results can be explained entirely with dataset-database biases, so this algorithm gives no support for the common assumption of a higher intrinsic OLG formation capacity of viruses compared with bacteria or eukaryotes. Two of the main differences between the taxonomic groups are the expected mutation rates and the average length of a protein. While genomes with higher mutation rates explore sequence space faster and therefore their proteins should appear to be more 'flexible' (i.e. less constrained), despite having the highest mutation rate virus domains do not actually appear to be very flexible. Another factor which deserves

further exploration is the age of a protein family, i.e. the time since gene birth. This may correlate with apparent 'sequence flexibility', which is the strongest influence on the result via the threshold values, due to increasing mutational saturation in older protein families. Being able to distinguish genuine sequence flexibility from mutational saturation, even in broad terms, could be very useful here. The length of the sequences on the other hand has been removed as a factor in this analysis. An artificial factor not considered could be database biases or an exchange matrix (BLOSUM62) biased towards certain kinds of proteins. The latter could be tested by using different matrices created from sequences from different taxonomic groups. It would be important to use the new matrix not only in the construction of the OLGs but also in the evaluation by the HMMs. So far it is not clear why protein families from different taxonomic groups are so different in their calculated ability to create OLGs.

The construction of overlapping genes also opens up many interesting possibilities for synthetic biology. For instance, mutations in overlapping regions are expected to be more deleterious on average, so an artificial genome with many OLGs is not only smaller but also expected to be more stable over time on a population level, as mutations are more likely to be strongly selected against. A recent method for stabilizing synthetic genes [67], where an arbitrary ORF was constructed to overlap a gene of interest and was concatenated with an essential gene downstream, could be taken a large step forward by overlapping whole genes thereby creating a system where not only 'polar' mutations are selected against but also more minor mutations, if they also affect the mother gene. Genome size has become a significant limiting factor for biomolecular computing, in which genetic programs are inserted into cells [68]. Existing compression methods [69] could be greatly improved by using OLGs, making more complex systems possible. In this context a well-designed stable synthetic genome could include fail-safe measures, such that faulty genetic programs would shut down. In summary, a better theoretical understanding of overlapping genes will be extremely useful in microbial genome annotation methods, the study of evolution, and in synthetic biology, and therefore deserves renewed attention.

Conclusions

In this study we have shown that encoding overlapping protein domains is possible for a large majority of arbitrary domain pairs, as assessed by Hidden Markov Models based on Pfam protein domain families. Similar secondary structure to natural proteins was also able to be achieved in constructed sequences. Contrary to previous reports, viral proteins were not easier to overlap than those from other taxonomic groups. The success rate in different reading frames however matches expectations based on combinatorial constraints, validating previous key theoretical work on overlapping genes in different relative reading frames. This research helps in understanding the ubiquity of overlapping genes (OLGs) across the domains of life, implies that OLGs should be sought outside of virus genomes, and supports the potential use of synthetically constructed overlapping genes in diverse areas of biotechnology. Additional research into the theoretical underpinnings of natural overlapping gene pairs building on this work will further improve our understanding of molecular evolution and genome annotation, and will provide new opportunities for synthetic biology.

Methods

Applying Hidden Markov Models for OLG evaluation

Pfam consists of a 'seed' database containing trusted sequences for each protein domain family as well as a 'full' database containing all the sequences of Uni-Prot sorted into the different families using HMM profiles created from the 'seed' database. Here the 'seed' database is used to create HMM profiles using HMMER3 (v3.2.1) [70] and the 'full' database is tested against those profiles, in order to find the representative sequence (that which scores highest) and to calculate the naturally occurring deviations. The representative sequence chosen for OLG construction and to determine the natural variation in homologues should be the most typical sequence in the respective protein domain. Choosing a randomly picked sequence instead of a representative one could result in choosing an outlier, resulting in unreliable comparison scores and less chance of designed sequences retaining functionality. While it is not clear how to determine the most typical sequence using BLAST, it is straightforward using Hidden Markov Models (HMMs), which reflect the 'average' sequence and are therefore a good representative for the whole protein family. When constructing HMM profiles from trusted homologues of a specific protein domain, the sequence with the best fit to the profile is chosen for OLG construction. The scores of the remaining sequences define the typical deviations of the protein domain and are used to define family-specific thresholds for the constructed sequences.

Choosing a family-specific threshold value takes care of most of the length dependencies, but in order to be sure and to be able to compare sequences of different lengths, each score resulting from a comparison between a sequence and a HMM profile is divided by the sequence length. Here scores are used instead of e-values, as the latter also depend on the database size, an arbitrary factor in this analysis. Aligning the best sequence with the 'seed' sequences using MAFFT (v7.419) [71], weights

used for sequence construction can be determined just as in [42]. A more detailed description of the calculation of the weights and their influence can be found below. When studying the influence of a protein family's taxonomic classification on the construction of OLGs, the 'seed' and the 'full' database are first filtered by the four major taxonomic groups - archaea, bacteria, eukaryotes and viruses - before creating the profiles and the thresholds. MUSCLE (v3.8.31) [72] was used for realigning the 'seed' sequences after taxonomic filtering. For subsequent analyses, random sets from the ~17,000 Pfam families were chosen, with the condition that each family must have at least 10 'seed' sequences and 4 'full' sequences in order for the weights and the thresholds to be reasonably defined. Each dataset consists of 150 families since the variance of the resulting OLG success rate barely declines for larger sets (see supplementary Fig. S9). Figure 10 summarizes the workflow.

Determining the average success rate from random overlap positions

In order to estimate the expected success rate of an individual overlap attempt, the domains are overlapped such that one domain is fully embedded into the other, at a random position. Just as in [42] the sequence with the lower quality of the two constructed OLGs is used as a conservative representative of the pair. After determining the success for each position, the percentage of successful positions for each OLG pair, the average success rate in each reading frame, and the overall success rate averaged across reading frames were calculated. The number of possible positions for each OLG pair is equal to their difference in length plus one, so using more than one overlap position in each pair is only possible for genes with different lengths. Increasing the number of positions with overlaps calculated for each gene does not change the expected success rate but reduces its variation between different sets (see supplementary Fig. S10). Comparing the variation caused by choosing random positions and the variation caused by choosing random Pfam families, the former turns out to be negligible and consequently only a single randomly chosen position for each OLG pair is used for subsequent analyses.

Length dependence of the HMM evaluation

In order to determine whether the relative evaluation of OLGs really removed the length dependency, the average quality 'Q' of an OLG pair is determined and compared for OLG pairs with different lengths. Q is defined as the ratio of the scores of the constructed sequence (S) over the original sequence (S_max) times 100. The quality is therefore the percentage score loss due to the overlap. Supplementary Fig. S11 shows the mean quality for datasets with different sequence lengths. Starting from around 50 amino acids, Q is indeed mostly independent of sequence length. The low Q values of smaller

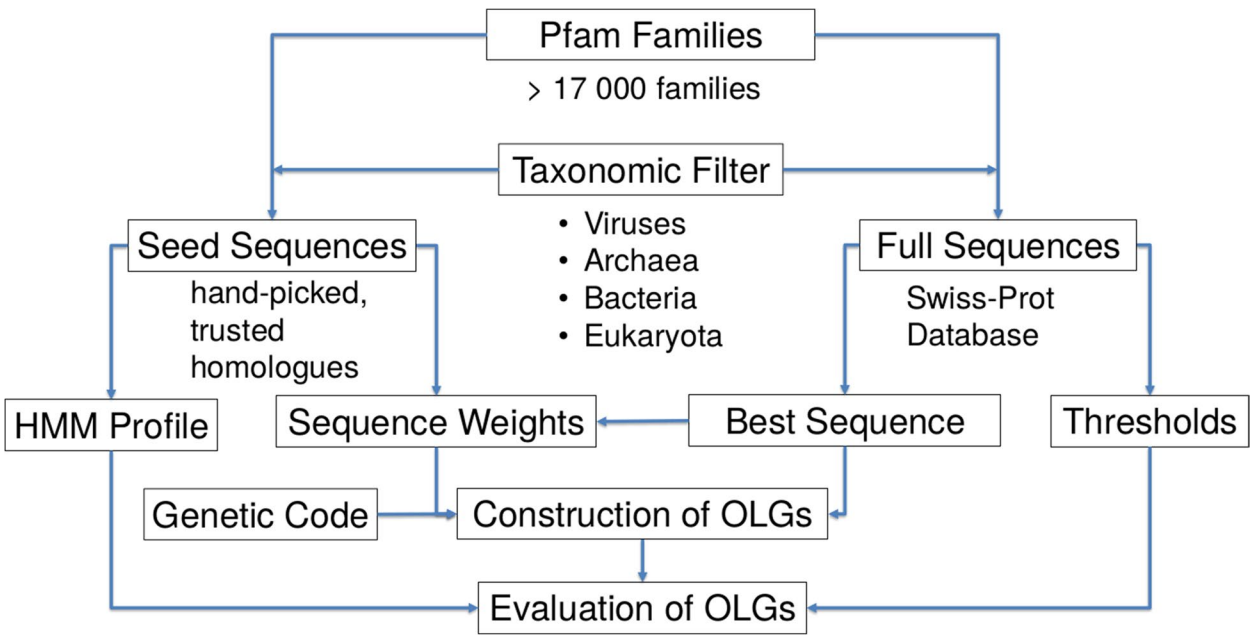

Fig. 10 Workflow for OLG construction and evaluation using HMMs and the Pfam database. HMM profiles are constructed from the seed sequences. The sequence with the highest score from the full group is used for OLG design. The remaining sequences in the full group are used to construct threshold scores used to evaluate the designed OLGs

sequences are because these sequences are less frequently matched to their respective HMM-profile, which results in a score of zero. The reason is probably that the shorter sequences fall below internal detection thresholds of HMMER3 more easily. Changing a single amino acid in a short gene changes its quality to a greater extent than in a long gene, resulting in larger fluctuations, which can lower the sequence below detection thresholds. Lowering internal thresholds of HMMER3 did not lead to more sequences being recognized by their respective profile.

In further analysis the minimum sequence length of 70 amino acids is used so that the percentage of OLG pairs in which at least one sequence is not recognised is below 5% (see Supplementary Fig. S11). When taking both sequences of each pair and not only the one with the lower quality, the quality distribution converges to a broad peak at around 76% with increasing sequence length (see Supplementary Fig. S12). Since the quality also depends on the flexibility of the HMM profiles used to score the sequences the peak is not expected to get any narrower with increasing sequence length and thus to reduce variations in sequence similarities between the constructed and the original sequences.

Optimisation of strength of positional weighting

The algorithm to construct OLG sequences from Opuu et al. uses an exchange matrix (Blosum62 [73]) to find the closest overlapping sequences to the original ones. It determines the codon with the highest sum of the scores for the exchanges in both sequences at each position. Sequence weights can prioritise the score of either one or the other sequence at different positions in order to increase the chance of obtaining functional sequences. In [42], the weight w_i at position i of the sequence is w_i=e^(-S_i), where S_i is the entropy calculated at position i in the alignment. The weights could be defined differently such that their influence on OLG construction is stronger or weaker. In order to optimize the weight strength a factor k is added to the entropy in their calculation such that w_i=e^(-kS_i). Varying k>0, the optimal weight strength for constructing OLGs can be determined, while k=0 means no weights are being used. In the HMM evaluation the influence of k is very weak. A value of k=0.5 is used in order to maximise the quality, Q (see Supplementary Fig. S13). Picking very high k values Q goes to zero. In this case at each position the sequence with the higher conservation maintains its amino acid. This indicates that it is crucial that at each position both sequences are changed in order to create functional OLGs.

In the BLAST evaluation k=0 is optimal (see Supplementary Fig. S14), such that no better value can be found for k>0. BLAST does not take special account of conserved regions of a sequence, so weights can improve one sequence but at the same time will reduce the score of the other. Since the lowest scoring of the two sequences is taken to represent the OLG pair, introducing weights has a high chance of reducing the success rate in an evaluation using BLAST, despite increasing biological relevance. This makes an evaluation using HMM or any other method that takes into account sequence conservation significantly preferable for judging constructed OLG pairs.

Abbreviations

OLG: overlapping gene; ORF: open reading frame; SGC: standard genetic code; HMM: hidden Markov model.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-08181-1.

Additional file 1.

Acknowledgements

Not applicable.

Authors' contributions

SW analyzed the data, developed the methods, and wrote the first draft. SS assisted with editing the manuscript and planning the study. ZA conceived of and supervised the study and assisted in drafting the manuscript. The author(s) read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. No additional funding was acquired for this research.

Availability of data and materials

The datasets created and analysed the current study, and associated Python scripts, are available from the corresponding author on reasonable request. The protein domain families are from Pfam version 32.0, http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/ and Swissprot version 2018_07.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Chair of Microbial Ecology, Department of Molecular Life Sciences, Technical University of Munich, Freising, Germany. [2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK.

Received: 8 November 2020 Accepted: 17 November 2021
Published online: 11 December 2021

References

1. Barrell BG, Air GM, Hutchison CA. Overlapping genes in bacteriophage φX174. Nature. 1976;264:34–41. doi:https://doi.org/10.1038/264034a0.

2. Cassan E, Arigon-Chifolleau A-M, Mesnard J-M, Gross A, Gascuel O. Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. Proc Natl Acad Sci U S A. 2016;113:11537–42. doi:https://doi.org/10.1073/pnas.1605739113.
3. Gelsinger DR, Dallon E, Reddy R, Mohammad F, Buskirk AR, DiRuggiero J. Ribosome profiling in archaea reveals leaderless translation, novel translational initiation sites, and ribosome pausing at single codon resolution. Nucleic Acids Res. 2020;48:5201–16. doi:https://doi.org/10.1093/nar/gkaa304.
4. Zehentner B, Ardern Z, Kreitmeier M, Scherer S, Neuhaus K. A Novel pH-Regulated, Unusual 603 bp Overlapping Protein Coding Gene pop Is Encoded Antisense to ompA in Escherichia coli O157:H7 (EHEC). Front Microbiol. 2020;11:377. doi:https://doi.org/10.3389/fmicb.2020.00377.
5. 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;8:17875. doi:https://doi.org/10.1038/s41598-018-35756-y.
6. 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:https://doi.org/10.3389/fmicb.2018.00931.
7. Loughran G, Zhdanov AV, Mikhaylova MS, Rozov FN, Datskevich PN, Kovalchuk SI, et al. Unprecedentedly efficient CUG initiation of an overlapping reading frame inPOLGmRNA yields novel protein POLGARF. doi:https://doi.org/10.1101/2020.03.06.980391.
8. Khan YA, Jungreis I, Wright JC, Mudge JM, Choudhary JS, Firth AE, et al. Evidence for a novel overlapping coding sequence in POLG initiated at a CUG start codon. BMC Genet. 2020;21:25. doi:https://doi.org/10.1186/s12863-020-0828-7.
9. Lèbre S, Gascuel O. The combinatorics of overlapping genes. J Theor Biol. 2017;415:90–101. doi:https://doi.org/10.1016/j.jtbi.2016.09.018.
10. Yockey HP. Do overlapping genes violate molecular biology and the theory of evolution? J Theor Biol. 1979;80:21–6. doi:https://doi.org/10.1016/0022-5193(79)90176-0.
11. Kolata GB. Overlapping genes: more than anomalies? Science. 1977;196:1187–8. doi:https://doi.org/10.1126/science.196.4295.1187.
12. Warren AS, Archuleta J, Feng W-C, Setubal JC. Missing genes in the annotation of prokaryotic genomes. BMC Bioinformatics. 2010;11:131. doi:https://doi.org/10.1186/1471-2105-11-131.
13. NCBI Prokaryotic Genome Annotation Standards. https://www.ncbi.nlm.nih.gov/genome/annotation_prok/standards/. Accessed 2 Oct 2021.
14. Dinan AM, Lukhovitskaya NI, Olendraite I, Firth AE. A case for a negative-strand coding sequence in a group of positive-sense RNA viruses. Virus Evol. 2020;6:veaa007. doi:https://doi.org/10.1093/ve/veaa007.
15. Nelson CW, Ardern Z, Goldberg TL, Meng C, Kuo C-H, Ludwig C, et al. Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic. Elife. 2020;9. doi:https://doi.org/10.7554/eLife.59633.
16. Meydan S, Vázquez-Laslop N, Mankin AS. Genes within Genes in Bacterial Genomes. Microbiol Spectr. 2018;6. doi:https://doi.org/10.1128/microbiolspec.RWR-0020-2018.
17. Ardern Z, Neuhaus K, Scherer S. Are Antisense Proteins in Prokaryotes Functional? Front Mol Biosci. 2020;7:187. doi:https://doi.org/10.3389/fmolb.2020.00187.
18. Belshaw R, Pybus OG, Rambaut A. The evolution of genome compression and genomic novelty in RNA viruses. Genome Res. 2007;17:1496–504. doi:https://doi.org/10.1101/gr.6305707.
19. Brandes N, Linial M. Gene overlapping and size constraints in the viral world. Biol Direct. 2016;11:26. doi:https://doi.org/10.1186/s13062-016-0128-3.
20. Scherbakov DV, Garber MB. Overlapping genes in bacterial and phage genomes. Mol Biol. 2000;34:485–95. doi:https://doi.org/10.1007/bf02759558.
21. Sáenz-Lahoya S, Bitarte N, García B, Burgui S, Vergara-Irigaray M, Valle J, et al. Noncontiguous operon is a genetic organization for coordinating bacterial gene expression. Proc Natl Acad Sci U S A. 2019;116:1733–8. doi:https://doi.org/10.1073/pnas.1812746116.
22. Ohno S. Evolution by Gene Duplication. Springer Berlin Heidelberg; 2014. https://play.google.com/store/books/details?id=0CbMoQEACAAJ.
23. Keese PK, Gibbs A. Origins of genes:" big bang" or continuous creation? of the National Academy of Sciences. 1992. https://www.pnas.org/content/89/20/9489.short.
24. Zull JE, Smith SK. Is genetic code redundancy related to retention of structural information in both DNA strands? Trends Biochem Sci. 1990;15:257–61. doi:https://doi.org/10.1016/0968-0004(90)90048-g.
25. Blalock JE, Others. Complementarity of peptides specified by'sense'and'antisense'strands of DNA. Trends Biotechnol. 1990;8:140–4. https://www.cabdirect.org/cabdirect/abstract/19901615648.
26. Štambuk N, Konjevoda P, Turčić P, Kövér K, Kujundžić RN, Manojlović Z, et al. Genetic coding algorithm for sense and antisense peptide interactions. Biosystems. 2018;164:199–216. doi:https://doi.org/10.1016/j.biosystems.2017.10.009.
27. Bartonek L, Braun D, Zagrovic B. Frameshifting preserves key physicochemical properties of proteins. Proc Natl Acad Sci U S A. 2020;117:5907–12. doi:https://doi.org/10.1073/pnas.1911203117.
28. Xu H, Zhang J. On the Origin of Frameshift-Robustness of the Standard Genetic Code. Mol Biol Evol. 2021;38:4301–9. doi:https://doi.org/10.1093/molbev/msab164.
29. Pavesi A, Magiorkinis G, Karlin DG. Viral Proteins Originated De Novo by Overprinting Can Be Identified by Codon Usage: Application to the "Gene Nursery" of Deltaretroviruses. PLoS Computational Biology. 2013;9:e1003162. doi:https://doi.org/10.1371/journal.pcbi.1003162.
30. Willis S, Masel J. Gene Birth Contributes to Structural Disorder Encoded by Overlapping Genes. Genetics. 2018;210:303–13. doi:https://doi.org/10.1534/genetics.118.301249.
31. Sabath N, Wagner A, Karlin D. Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol. 2012;29:3767–80. doi:https://doi.org/10.1093/molbev/mss179.
32. Carter CW Jr. Simultaneous codon usage, the origin of the proteome, and the emergence of de-novo proteins. Curr Opin Struct Biol. 2021;68:142–8. doi:https://doi.org/10.1016/j.sbi.2021.01.004.
33. Schlub TE, Holmes EC. Properties and abundance of overlapping genes in viruses. Virus Evolution. 2020;6. doi:https://doi.org/10.1093/ve/veaa009.
34. Pham Y, Li L, Kim A, Erdogan O, Weinreb V, Butterfoss GL, et al. A minimal TrpRS catalytic domain supports sense/antisense ancestry of class I and II aminoacyl-tRNA synthetases. Mol Cell. 2007;25:851–62. doi:https://doi.org/10.1016/j.molcel.2007.02.010.
35. Rodin SN, Ohno S. Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. Orig Life Evol Biosph. 1995;25:565–89. doi:https://doi.org/10.1007/BF01582025.
36. Carter CW, Popinga A, Bouckaert R, Wills PR. Class I Aminoacyl-tRNA Synthetase Urzyme and CP1 Modules have Distinct Genetic Origins. bioRxiv. 2020;:2020.04.09.033712. doi:https://doi.org/10.1101/2020.04.09.033712.
37. Martinez-Rodriguez L, Erdogan O, Jimenez-Rodriguez M, Gonzalez-Rivera K, Williams T, Li L, et al. Functional Class I and II amino acid-activating enzymes can be coded by opposite strands of the same gene. Journal of Biological Chemistry. 2016;291:23830–1. doi:https://doi.org/10.1074/jbc.a115.642876.
38. Carter CW, Li L, Weinreb V, Collier M, Gonzalez-Rivera K, Jimenez-Rodriguez M, et al. The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: an unlikely scenario for the origins of translation that will not be dismissed. Biology Direct. 2014;9. doi:https://doi.org/10.1186/1745-6150-9-11.
39. LéJohn HB, Cameron LE, Yang B, Rennie SL. Molecular characterization of an NAD-specific glutamate dehydrogenase gene inducible by L-glutamine. Antisense gene pair arrangement with L-glutamine-inducible heat shock 70-like protein gene. Journal of Biological Chemistry. 1994;269:4523–31. doi:https://doi.org/10.1016/s0021-9258(17)41809-6.
40. LéJohn HB, Cameron LE, Yang B, MacBeath G, Barker DS, Williams SA. Cloning and analysis of a constitutive heat shock (cognate) protein 70 gene inducible by L-glutamine. J Biol Chem. 1994;269:4513–22. doi:https://doi.org/10.1016/S0021-9258(17)41808-4.
41. Williams TA, Wolfe KH, Fares MA. No Rosetta Stone for a Sense–Antisense Origin of Aminoacyl tRNA Synthetase Classes. Mol Biol Evol. 2008;26:445–50. doi:https://doi.org/10.1093/molbev/msn267.
42. Opuu V, Silvert M, Simonson T. Computational design of fully overlapping coding schemes for protein pairs and triplets. Sci Rep. 2017;7:15873. doi:https://doi.org/10.1038/s41598-017-16221-8.

43. Wang B, Papamichail D, Mueller S, Skiena S. Two Proteins for the Price of One: The Design of Maximally Compressed Coding Sequences. DNA Computing. 2006;:387–98. doi:https://doi.org/10.1007/11753681_31.
44. Inouye M, Ishida Y, Inouye K. Designing of a single gene encoding four functional proteins. J Theor Biol. 2017;419:266–8. doi:https://doi.org/10.1016/j.jtbi.2017.01.042.
45. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical Assessment of Methods of Protein Structure Prediction (CASP) – Round XIV. Proteins: Structure, Function, and Bioinformatics. 2021. doi:https://doi.org/10.1002/prot.26237.
46. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9. doi:https://doi.org/10.1038/s41586-021-03819-2.
47. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins: Structure, Function, and Bioinformatics. 2019;87:1011–20. doi:https://doi.org/10.1002/prot.25823.
48. Blazejewski T, Ho HI, Wang HH. Synthetic sequence entanglement augments stability and containment of genetic information in cells. Science. 2019;365. doi:https://doi.org/10.1126/science.aav5477.
49. Fernandes JD, Faust TB, Strauli NB, Smith C, Crosby DC, Nakamura RL, et al. Functional Segregation of Overlapping Genes in HIV. Cell. 2016;167:1762–73.e12. doi:https://doi.org/10.1016/j.cell.2016.11.031.
50. Safari M, Jayaraman B, Yang S, Smith C, Fernandes JD, Frankel AD. Functional and Structural Segregation of Overlapping Helices in HIV-1. bioRxiv. 2021;:2021.07.15.452440. doi:https://doi.org/10.1101/2021.07.15.452440.
51. Wichmann S, Ardern Z. Optimality in the standard genetic code is robust with respect to comparison code sets. Biosystems. 2019;185:104023. doi:https://doi.org/10.1016/j.biosystems.2019.104023.
52. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 1997;28:405–20. doi:10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l.
53. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999;12:85–94. doi:https://doi.org/10.1093/protein/12.2.85.
54. Torrisi M, Kaleel M, Pollastri G. Deeper Profiles and Cascaded Recurrent and Convolutional Neural Networks for state-of-the-art Protein Secondary Structure Prediction. Sci Rep. 2019;9:12374. doi:https://doi.org/10.1038/s41598-019-48786-x.
55. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–637. doi:https://doi.org/10.1002/bip.360221211.
56. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, et al. A series of PDB related databases for everyday needs. Nucleic Acids Res. 2011;39 Database issue:D411–9. doi:https://doi.org/10.1093/nar/gkq1105.
57. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A. 1951;37:205–11. doi:https://doi.org/10.1073/pnas.37.4.205.
58. Freeland SJ, Hurst LD. The genetic code is one in a million. J Mol Evol. 1998;47:238–48. doi:https://doi.org/10.1007/pl00006381.
59. Massey SE. A neutral origin for error minimization in the genetic code. J Mol Evol. 2008;67:510–6. doi:https://doi.org/10.1007/s00239-008-9167-4.
60. Zhao F, Yu C-H, Liu Y. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. Nucleic Acids Res. 2017;45:8484–92. doi:https://doi.org/10.1093/nar/gkx501.
61. Miyata T, Yasunaga T. Evolution of overlapping genes. Nature. 1978;272:532–5. doi:https://doi.org/10.1038/272532a0.
62. Butler T, Goldenfeld N, Mathew D, Luthey-Schulten Z. Extreme genetic code optimality from a molecular dynamics calculation of amino acid polar requirement. Phys Rev E Stat Nonlin Soft Matter Phys. 2009;79 6 Pt 1:060901. doi:https://doi.org/10.1103/PhysRevE.79.060901.
63. Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, et al. PredictProtein—an open resource for online prediction of protein structural and functional features. Nucleic Acids Research. 2014;42:W337–43. doi:https://doi.org/10.1093/nar/gku366.
64. Bartonek L, Zagrovic B. VOLPES: an interactive web-based tool for visualizing and comparing physicochemical properties of biological sequences. Nucleic Acids Res. 2019;47:W632–5. doi:https://doi.org/10.1093/nar/gkz407.
65. Tian P, Best RB. How Many Protein Sequences Fold to a Given Structure? A Coevolutionary Analysis. Biophys J. 2017;113:1719–30. doi:https://doi.org/10.1016/j.bpj.2017.08.039.
66. Tian P, Louis JM, Baber JL, Aniana A, Best RB. Co-evolutionary fitness landscapes for sequence design. Angew Chem Int Ed Engl. 2018;57:5674–8. doi:https://doi.org/10.1002/anie.201713220.
67. Decrulle AL, Frenoy A, Meiller-Legrand TA, Bernheim A, Lotton C, Gutierrez A, et al. Engineering gene overlaps to sustain genetic constructs in vivo. doi:https://doi.org/10.1101/659243.
68. Benenson Y. Biomolecular computing systems: principles, progress and potential. Nat Rev Genet. 2012;13:455–68. doi:https://doi.org/10.1038/nrg3197.
69. Lapique N, Benenson Y. Genetic programs can be compressed and autonomously decompressed in live cells. Nat Nanotechnol. 2018;13:309–15. doi:https://doi.org/10.1038/s41565-017-0004-z.
70. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14:755–63. doi:https://doi.org/10.1093/bioinformatics/14.9.755.
71. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80. doi:https://doi.org/10.1093/molbev/mst010.
72. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7. doi:https://doi.org/10.1093/nar/gkh340.
73. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89:10915–9. doi:https://doi.org/10.1073/pnas.89.22.10915.

Publisher's Note