Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Evolutionary selective constraints acting on the stop codon across land plants

Vishvesh Karthik [1],* and Cathal Seoighe [2],*

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: All genomes are under an evolutionary pressure and struggle to keep the functional portion of the DNA. The rate at which favorable genes are retained and deleterious ones are lost is exerted by a parameter which is the ratio of synonymous(dS) to non synonymous(dN) mutation rates. Substitutions do not alter the coded amino acid while mutations do. This makes substitutions helpful and mutations harmful. When $\frac{dN}{dS} < 1$, substitution rate is greater than mutation rate and the gene is said to be under purifying selection. Purifying selection favors synonymous substitutions than non-synonymous mutations thereby preventing change of an amino acid residue at a give position. In conventional models of substitution only the sense(non-stop) codons are accounted for while the non-sense codons are omitted because they do not contribute to amino acid changes. Since stop codons function with varying efficiencies, they can be read-through and have the ability to alter the final protein products. When combined with other mechanisms like ribosome stalling and mRNA regulation, stop codons can indirectly modulate protein synthesis. This gives meaning to stop codon preservation and substitution, thereby creating the need to include them in standard models of substitution. Stop codons have a low probability of undergoing mutations [1] but the pressure acting on their rate of substitution is only vaguely addressed. *Seioghe et al.* have introduced a new model which incorporates stop codons into the general Muse & Gaut substitution model [2]. The model is constructed based on the assumption that stop codons are also under selection pressure and co-evolve with the genes. The extended Muse & Gaut model, casually called extMG model, has been applied on mammalian orthologous sequences and 50% of the genes were found to be under purifying selection. In this study the extMG model of substitution, for all 64 codons, is used to estimate ϕ (rate of substitution between stop codons) for plant ortholog families under the *Viridiplantae* clade.
Results: The mixture weight informs that 60% of all the stop codons analyzed in the plant genes are under purifying selection. Out of all the stop codons half of UGA stop codons show preservation, while 70% of UAG and 60% of UAAs are more likely to switch to UAA or UGA/UAG respectively.
Availability: All the code is available on github and is open access ($https : //github.com/vizkidd/stop\_codon\_plants$)
Contact: vishveshkarthik@gmail.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

After the transcription of DNA to mRNA, ribosomes populate the mRNA and start translating it to proteins. Each translating ribosome reads nucleotides in frames of three, called codons, and builds a corresponding amino acid chain. When an mRNA associated with a gene is translated into protein it is termed as expression. A series of sense codons are translated until a stop codon is reached. Stop codons, termed non-sense codons, mark the end of translation. Stop codons have different efficiencies ($UAA > UAG > UGA$) in terminating translation. Recent developments in literature suggest that ribosomes can read-through stop codons for the regulation of protein synthesis [3]. This mechanism uses readthroughs in human AMD1 mRNAs to skip the primary stop codon and park ribosomes at a stop codon in the 3' UTR region. The length of the ribosomal queue

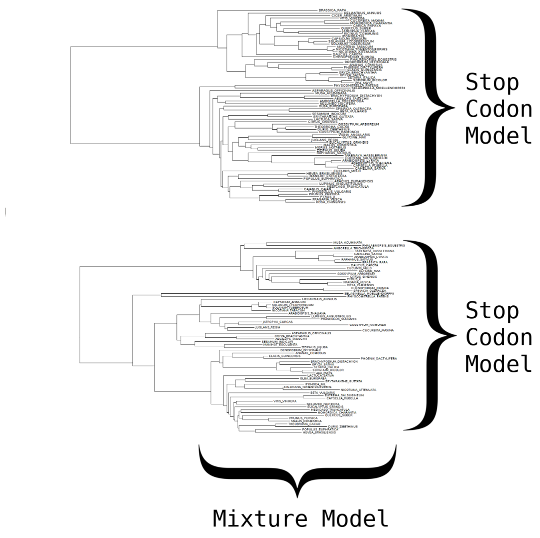

Fig. 1. Application of Stop Codon Model & Mixture Model on the latitudinal and longitudinal data of OGs.

is proportional to the proteins synthesized. Analysis of multiple taxa has shown that overall stop codon abundance in the genome is associated with shorter protein sequences [4]. Frequent occurances of stop codons in the sequences would neither support readthroughs nor longer deleterious C-terminus extensions in proteins. Stop codon efficiencies play an important role in modulating such regulatory and mRNA quality control mechanisms which help to maintain flexible gene expression. An analysis for examining selective constraints acting on stop codons was performed on mammalian protein coding sequences and is based on the preamble that stop codons are non-conformant to random mutations but are under selective pressure. This supports the axiom that stop codons co-evolve with the genes of which they are a part of and are under constant selective pressure. The study of stop codons in mammalian CDS uncovered that 57% of the genes were under purifying selection, especially the UGA stop codon which is observed in nearly 50% of the mammalian genes [5]. Orthologous gene groups, which are the most informative about conservation, selection and evolution of the same gene across different species is used for the analysis. For this study, the same evolutionary phylogenetic analysis pipeline is implemented on plant sequences. Similar analysis for estimating pressure of selection on stop codons have been implemented at smaller scales with few orthologous genes in a small groups of prokaryotic organisms [6] & humans [7]. The scale of such analysis do not have the necessary power to derive the rate at which stop codons are substituted. Our analysis implements a novel codon substitution model which includes stop codons under the hypothesis that stop codons are co-evolving with the genes themselves. Estimation of selection in stop codons is valuable information in determining the true rate of evolution, to isolate proteins under readthrough contexts and the significance of genes which code for such proteins.

2 Approach

The analysis is focused on orthologous genes across plant species. A list of orthologous cluster groups are required for the analysis. Several databases such as Phytozome, PLAZA, PlantOrDB, POGs2 were accessed [8] but IC4R and OrthoDB [9] were selected from the list because of the availability of plant ortholog clusters based on gene families. The latter of the two is preferred for two reasons,

- Orthologs are more explanatory in terms of evolution of the gene from a common ancestor and doesn't contain information on duplicated genes unlike paralogs
- OrthoDB contains more orthologous groups(OGs) from a variety of species when compared to IC4R (which is a homolog database restricted to sub-species of rice).

Table 1. Extended MG model parameters

Parameters	Description
κ	T$_s/T_v ratio$
ω	dN/dS ratio
s	Scale parameter for branch lengths
ϕ	SC substitution rate relative to synonymous substitution rate

SC-Stop Codon

Organisms under a specific taxonomic level are isolated and their parent OGs are obtained. These subset of OGs are trimmed to isolate groups based on a higher taxonomic level. This is done to segregate OGs based on a higher clade and organisms based on a lower clade/family. The protein coding sequences (CDS) for all the genes under each OG are downloaded from NCBI. These sequences are then refined to remove different transcripts from the same organism. The inframe-CDS sequences are translated to amino acid and an initial alignment is performed. The aligned amino acid sequences are clustered based on stop codons (casually called stop clusters or SC). Unaligned nucleotide CDS are segregated based on SCs (ntSC), translated to amino acid sequences and aligned. The general Muse & Gaut codon substitution model [2] is extended to include the three stop codons. This extended MG model is applied to the ntSCs to estimate the model parameters κ, ω, ϕ & s (scaling factor for branch lengths). The rate matrix for the extended model consists of:

$$\text{q}_{ij} = \begin{cases} \pi_{jk} \text{ , synonymous transversion } (i, j \in S) \\ \kappa\pi_{jk} \text{ , synonymous transition } (i, j \in S) \\ \omega\pi_{jk} \text{ , non-synonymous transversion } (i, j \in S) \\ \kappa\omega\pi_{jk} \text{ , non-synonymous transition } (i, j \in S) \\ 0 \text{ , } > \text{ 1 nucleotide difference} \\ \phi\kappa\pi_{jk} \text{ , synonymous transition } (i, j \in N) \\ 0 \text{ , i } \in \text{N} \oplus \text{j} \in \text{ N} \end{cases} \tag{1}$$

Where q$_{ij}$ is the generator matrix, S and N are sets of sense and stop codons respectively. The extMG model assumes equal codon equilibrium frequencies $\pi_1 = \pi_2 = \pi_3$. This when combined with MG model is called MGF1x4 (One equilibrium frequency for all the codons in a frame). ϕ is used to model the rate of substitution between stop codons. Transversions are not considered because only transitions are possible between stop codons. Only instantaneous rates are used and dinucleotide substitutions have a rate of 0, therefore, it is not possible to traverse from every state to every other state [Fig. 2]. Thus the transition probability matrix is not irreducible and does not have a unique stationary distribution.

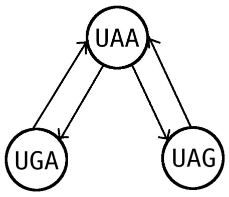

Fig. 2. Non-irreducible markov chain constructed with stop codons.

3 Methods

3.1 Cluster Selection

Flat files containing the species, taxonomic levels and OGs were downloaded from OrthoDB. Since the analysis was focused on land plants, the taxonomic division of *Embryophyta* (which has a NCBI taxonomic ID of 3193) was chosen. Organisms were extracted for the given taxonomy ID and a total of 40 organisms were selected. The ortholog clusters were then converted from gene-based clusters to organism-based clusters to reduce processing time and the clusters in which the organisms are a member of, are isolated. These clusters were filtered and ranked based on the number of participating organisms. For filtering, counts of organisms in clusters was limited from (2 x number of selected organisms) to (scale factor * selected organisms). A final selection process based on cluster level taxa ID confined the clusters to taxonomic levels which fall under the given taxonomic ID. This is a flexible way of choosing and analyzing clusters and determining the contribution of the selected organisms to the selected clusters. For this study the organisms which fall under the *Embryophyta* clade were chosen and the clusters under *Viridiplantae* to which the organisms contribute to were picked [Fig. 3]. Clusters, for which the unique organism count is greater than the count of selected organisms, were kept.

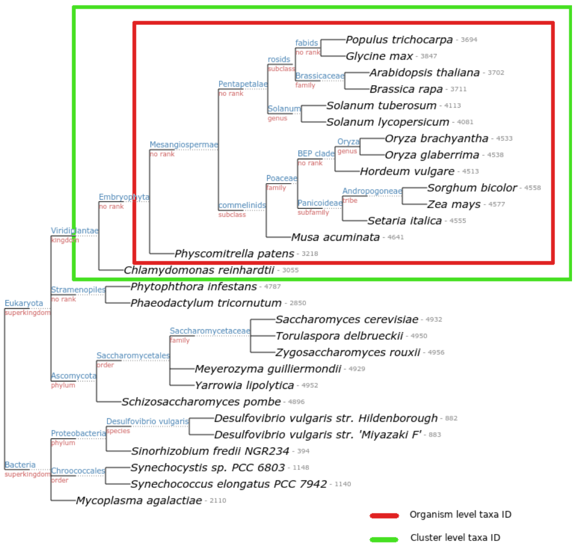

Fig. 3. Taxonomic level coverage of example orthologous groups

3.2 Sequence Extraction and Quality Control

All the genes in each of the clusters, for the selected organisms, were downloaded from NCBI after looking up the gene information using OrthoDB API. The quality of the sequences were checked and the sequences which were not in-frame or did not have a stop codon were discarded. Duplicates with the same gene IDs and empty sequences were deleted and the data was sorted. If an organism had more than one sequence in a cluster, only one of the transcripts was chosen by calculating the mean sequence length for all sequences. The sequence of which the length deviated the least from the mean was chosen. This process chooses the best transcript from each organism in a cluster thereby promoting good alignments.

3.3 Data Processing

The CDS sequences were translated to amino acids and the amino acid sequences were subject to an initial alignment(without refinements) using DECIPHER package in R [10]. The aligned nucleotide sequences were clustered based on stop codon locations and the corresponding unaligned NT sequences were segregated based on the stop codon positions(casually called stop clusters). This was done based on the assumption that the sequences for which the stop codons are at the same position were more closely related. The systematic bias which can be introduced by this forced grouping by stop codon positions is overruled when considering that the stop codons are under scrutiny. It also is an easy way to assimilate converging sequences and accounts for missing data to produce better alignments. Each of the stop clusters was then aligned using DECIPHER. The alignment was based on amino acid translations to maintain the frames of the codons. Extended MG (extMG) is applied to the NT alignments to estimate the model parameters and Nelder-Mead method [11] is used for optimization. A likelihood ratio test was performed between the NULL model with ϕ=1 and alternate model with ϕ as a free parameter, at a significance level of α=0.05 in order to test for the hypothesis that $\phi > \phi_0$ [Fig. 4]. Stop codons were trimmed from the alignments, converted to numbers and appended to a separate file. CodonPHYML [12] was used to infer trees and model parameters $(\kappa, \omega, \pi_A, \pi_G, \pi_C, \pi_T)$ for the stop codon free alignments. Finally, a Gaussian mixture model was applied on the longitudinal data. The mixture model has two point masses, one at $\phi = 1$ (which supports neutral evolution) and other point mass at $\phi < 1$ (supporting purifying selection). Positive selection is ignored because we are only interested in the genes which are being protected from deleterious mutation and purifying selection in general. The parameter estimates for ϕ & p were estimated using the mixture model and the values were tested for certainty using a bootstrapping method.

4 Results

Using the above mentioned pipeline, 1161 ortholog families, with an average of 26 organisms in each cluster, were downloaded. Out of 40 chosen organisms, 11 organisms were discarded since nucleotide

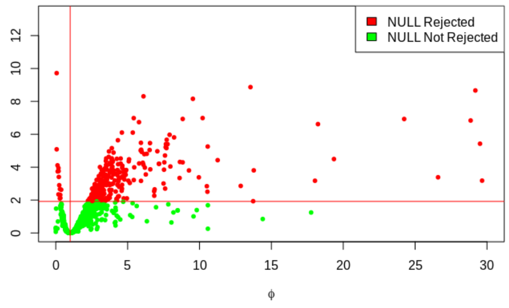

Fig. 4. χ^2 Test: H_0 :$\phi = 1$ & H_1 :$\phi > 1$ at a significance level of 0.05.

sequences for some protein coding genes were not available on GenBank [13]. The obtained sequences in each cluster aligned well with each other and no clusters sub-grouped based on the stop codon. Initial quality control of sequences and selection based on length deviation was sufficient to choose closely related sequences. Thus no bias was introduced by stop codon clustering. From the nucleotide sequences it can be inferred that UGA is the most abundant stop codon [Fig. 5]. UAG has low proportions and is sparely distributed while UAA has higher proportions and is more widely distributed [Fig. 7]. A χ^2 test is performed on the clusters at a significance level of $\alpha = 0.05$ and the NULL model of $\phi = 1$ is accepted for a large proportion of the clusters [Fig. 4]. The extMG model estimated that 686 of 1161 clusters are under neutral and purifying selection. The mixture model estimated the ϕ value as 0.3 and that 60% of the stop codons were under purifying selection. This mixture weight is very close to the prediction in mammals where 50% of the stop codons were predicted to be under purifying selection. However, bootstrap analysis produced less optimal results due to the small size of clusters and the low number of organisms participating in them. This caused the point masses to overlap with each other while estimating bootstrapping. The overlap created optima with minimum depth. This can be resolved by choosing clusters with better upper limits and including more genes under different selection pressures which create distinct masses.

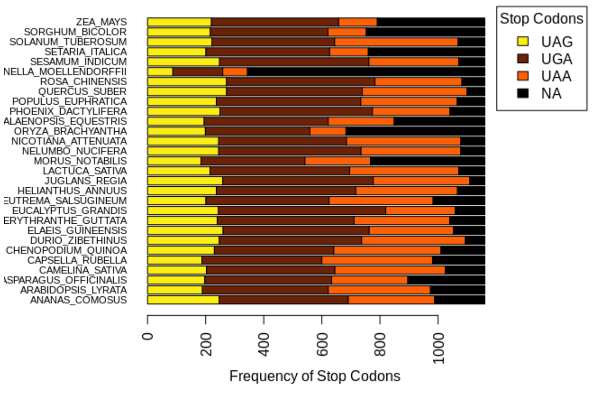

Fig. 5. Bar-plot which explains the frequency of all stop codons.

5 Discussion

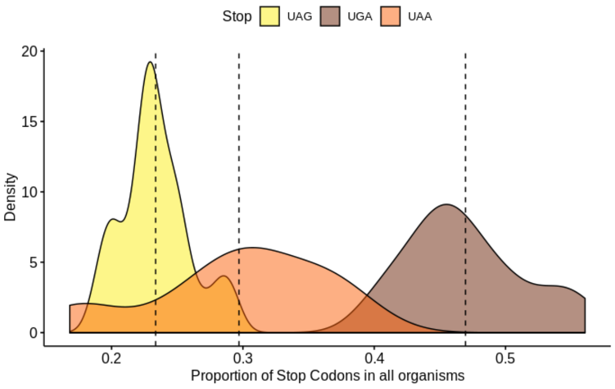

Fig. 6. Density plot of proportion of stop codons.

The estimated proportion of stop codons under purifying selection highlight that a lot of the stop codons are being preserved across plant species. This indicates that the stop codons have an additive effect in terms of evolution of the genes in the selected plant species. On manual examination of properties for genes which lie in the higher ϕ value range, we can see that many of the genes which substitute to keep the stop codons are part of key biological processes like DNA repair, defense mechanisms, cell-cycle, post translational modifications etc. UGA stop codon plays a minor role in coding for selenocysteine, which is a selenoprotein [14]. Selenoproteins are rare trace elements which play a vital role in host defense systems. Other novel regulatory mechanisms like auto-regulation of protein synthesis through read-throughs and parking of ribosomes in the 3'UTR region suggest that the necessity of the stop codons are mostly

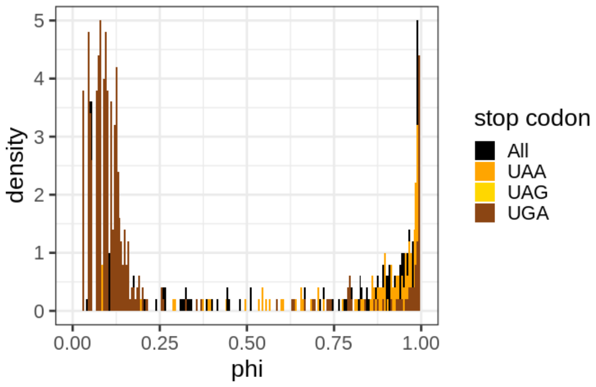

Fig. 7. Bootstrapped ϕ values for 1000 replicates.

Table 2. Stop codon properties [7]

Codon	Density	Frequency	GC-Relation	Efficiency	Preference
UAA	$>$	$\sim$	$<$	$>$	$>$
UGA	$\sim$	$>$	$>$	$<$	$\sim$
UAG	$<$	$<$	$\sim$	$\sim$	$<$

$>$-High
$\sim$-Intermediate
$<$-Low

under-estimated. The various properties and the general preference in the choice of stop codon provided in the table above [Table. 2] briefly points to the various variables which can be considered when selection acts for or against a stop codon. For estimating the selection pressure, a simple MG model was used due to it's simplicity, flexibility and negation of context dependent effects [15]. Though MGF1x4 model is not the best performing model [16], it can be further improved by the inclusion of codon preference statistics to improve overall performance and produce better prediction results. The markov chain constructed for the stop codons can be made irreducible by the addition of a fourth "NNN" state which will allow transitions from any state to any other state, this would frantically reduce the general 64x64 model into a 4x4 model focused only on stop codons but risks unwanted state transitions leading to mutations. This can be avoided by setting a low π and transition rates but the optimality of such a model is still questionable.

6 Conclusion

The new extMG model provides a chance to predict additive effects stop codons have in the true rate of evolution of genes. This information can be used to isolate genes which might be under readthrough contexts and would unlock more information as to the existence of those genes and such mechanisms. The rate of evolution of stop codons can also be used to predict whether a gene is being lost or gained and the phenotypes can be compared to evaluate external effects.

Acknowledgements

We are grateful to Cathal Seioghe, Anuprita Zore, Shweta Godbole and the anonymous reviewer, whose comments and insightful opinions greatly improved the manuscript.

References

[1] Tariq Abdullah, Muniba Faiza, Prashant Pant, Mohd Rayyan Akhtar, and Pratibha Pant. An analysis of single nucleotide substitution in genetic codons - probabilities and outcomes. *Bioinformation*, 12(3):98–104, June 2016.

[2] S V Muse and B S Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, 09 1994.

[3] Martina M. Yordanova, Gary Loughran, Alexander V. Zhdanov, Marco Mariotti, Stephen J. Kiniry, Patrick B. F. O'Connor, Dmitry E. Andreev, Ioanna Tzani, Paul Saffert, Audrey M. Michel, Vadim N. Gladyshev, Dmitry B. Papkovsky, John F. Atkins, and Pavel V. Baranov. Amd1 mrna employs ribosome stalling as a mechanism for molecular memory formation. *Nature*, 553:356, January 2018.

[4] Louise J. Johnson, James A. Cotton, Conrad P. Lichtenstein, Greg S. Elgar, Richard A. Nichols, p. David Polly, and Steven C. Le Comber. Stops making sense: translational trade-offs and stop codon reassignment. *BMC Evolutionary Biology*, 11(1):227, July 2011.

[5] Cathal Seoighe, Stephen J. Kiniry, Andrew Peters, Pavel V. Baranov, and Haixuan Yang. Selection shapes synonymous stop codon use in mammals. *bioRxiv*, 2019.

[6] Frida Belinky, Vladimir N. Babenko, Igor B. Rogozin, and Eugene V. Koonin. Purifying and positive selection in the evolution of stop codons. *Scientific Reports*, 8(1):9260, June 2018.

[7] Edoardo Trotta. Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. *BMC genomics*, 17:366–366, May 2016.

[8] Manuel Martinez. Computational tools for genomic studies in plants. *Current genomics*, 17(6):509–514, December 2016.

[9] Evgenia V. Kriventseva, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. Orthodb v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic acids research*, 47(D1):D807–D811, January 2019.

[10] Erik S. Wright. Using decipher v2.0 to analyze big biological sequence data in r. *The R Journal*, 8(1):352–359, 2016.

[11] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 01 1965.

[12] Manuel Gil, Marcelo Serrano Zanetti, Stefan Zoller, and Maria Anisimova. Codonphyml: fast maximum likelihood phylogeny estimation under codon substitution models. *Molecular biology and evolution*, 30(6):1270–1280, June 2013.

[13] Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. Genbank. *Nucleic acids research*, 44(D1):D67–D72, January 2016.

[14] Lian-Hai Fu, Xiao-Feng Wang, Yoram Eyal, Yi-Min She, Lynda J. Donald, Kenneth G. Standing, and Gozal Ben-Hayyim. A selenoprotein in the plant kingdom: Mass spectrometry confirms that an opal codon (uga) encodes selenocysteine in chlamydomonas reinhardtii glutathione peroxidase. *Journal of Biological Chemistry*, 277(29):25983–25991, 2002.

[15] Helen Lindsay, Von Bing Yap, Hua Ying, and Gavin A. Huttley. Pitfalls of the most commonly used models of context dependent substitution. *Biology direct*, 3:52–52, December 2008.

[16] Nicolas Rodrigue, Nicolas Lartillot, and Hervé Philippe. Bayesian comparisons of codon substitution models. *Genetics*, 180(3):1579–1591, November 2008.