Estimating the consequences of multiple simultaneous mutations on inferences of evolutionary selection

Alexander G. Lucaci; Sadie Wisotsky, PhD; Sergei L Kosakovsky Pond, PhD
Department of Biology, Temple University

Abstract

Contemporary codon models of molecular sequence evolution often assume changes within a specific codon happen stepwise as a single, instantaneous nucleotide substitution. However, these models may fail to accurately depict evolutionary pressures and constraints at individual sites along a protein-coding sequence or across branches. **Our novel statistical method modifies the MG94 (1) codon model to allow double and triple instantaneous substitutions within a codon in addition to the conventional single nucleotide replacement.** We apply our new codon model to simulated and empirical data sets to compare the results of Single, Double, and Triple nucleotide substitution inclusive models (SH, DH, TH). **Our goal is to determine the feasibility of modeling multiple simultaneous hits (MSH) within codons and if our ability to detect a signal is contingent upon realistic biological phenomenon or statistical noise.**

Background

Our codon model reflects that improvement in modeling and detection of natural selection in protein-coding sequences may require the inclusion of MSH's. Statistical model fits suggest that there seems to be a reliable MSH signal being measured. **In our empirical dataset (SELECTOME/Euteleostomi), we observed both DH and TH better fit the data for 94% and 93% of cases than SH, respectively. However, TH was preferred in only 37% of the data sets to DH.** While DH has biological support in the literature (2, 4), the biological backing for TH is rarer. Therefore, we sought to explore the magnitude and possible pluralistic nature of contributors to the TH signal. This is especially true in gene alignments where TH is preferred over DH and/or SH. Improved biological realism (5) within codon models may shed light on the additional routes available for genes and proteins to embark upon during adaptive evolutionary periods.

We define our Q matrix as follows:

- ω = dN/dS = β/α, the rate of nonsynonymous to synonymous substitutions
- $\square$ = double instantaneous mutation rate
- Ψ = triple instantaneous mutation rate
- Θ_{ij} = underlying nucleotide substitution rate (follows GTR form)
- π_j = nucleotide frequency of target nucleotide

Q_{ij}

$\alpha\Theta_{ij}\pi_j$	Synonymous	single nucleotide change
$\alpha\omega\Theta_{ij}\pi_j$	Nonsynonymous	single nucleotide change
$\alpha\delta\prod_{n=1}^{2}\theta^n_{ij}\pi^n_j$	Synonymous	double nucleotide change
$\alpha\omega\delta\prod_{n=1}^{2}\theta^n_{ij}\pi^n_j$	Nonsynonymous	double nucleotide change
$\alpha\Psi\prod_{n=1}^{3}\theta^n_{ij}\pi^n_j$	Synonymous	triple nucleotide change
$\alpha\omega\Psi\prod_{n=1}^{3}\theta^n_{ij}\pi^n_j$	Nonsynonymous	triple nucleotide change

Results

Figure 1. This plot demonstrates all triple instantaneous mutations analyzed from the **SELECTOME/Euteleostomi (dataset n=13,303).** The results below are filtered (>200), to show the most common **codon to codon exchanges (n=24,653)**

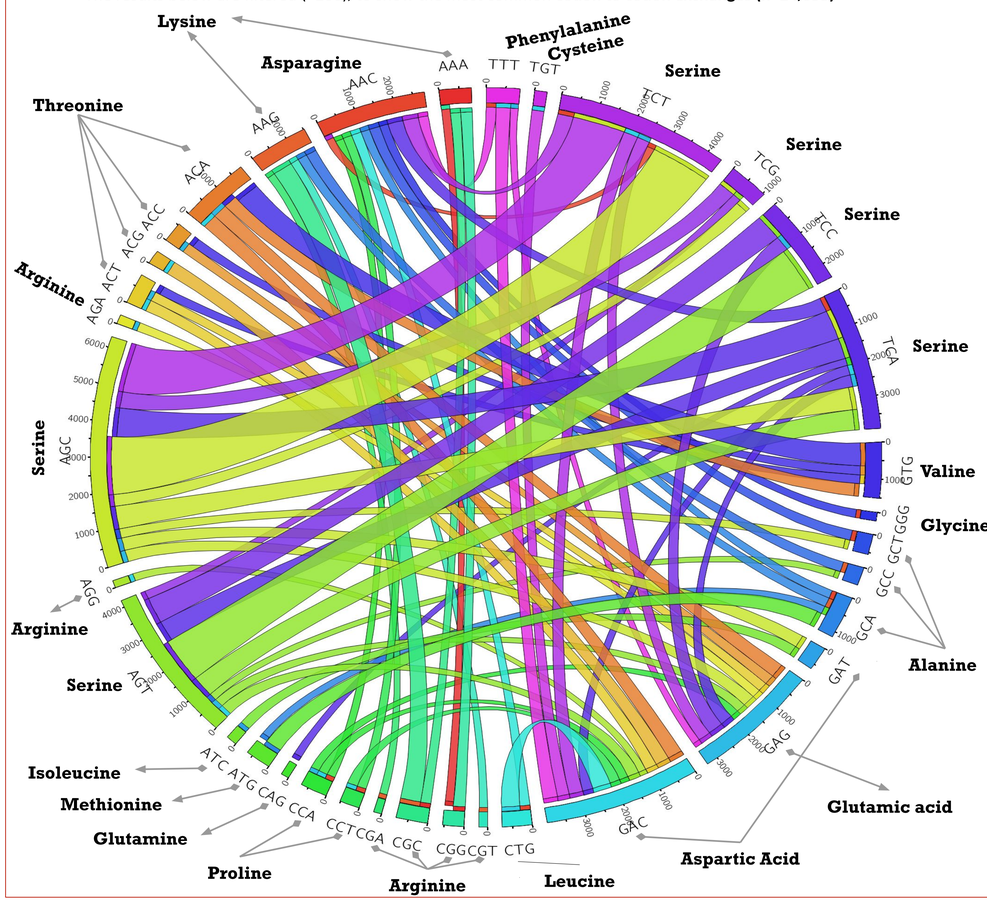

Figure 2. Comparison of omega values across rate categories (n=3), which are used to determine the type and scale of selection at individual sites.

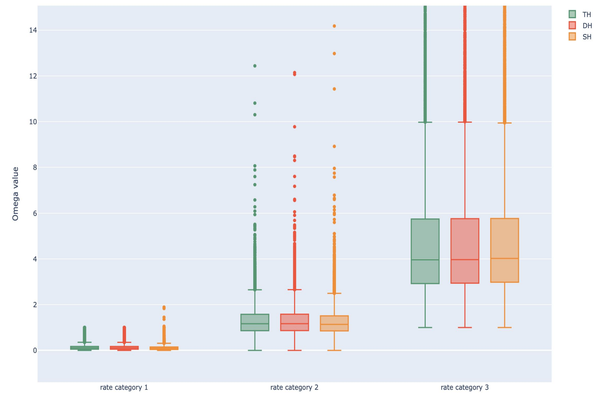

Figure 3. GO analysis for gene alignments where the TH model is preferred over the DH ($p < 0.05$). TH's may be a feature of specific protein classes which warrants further investigation.

	TH (TH + DH + SH inclusive)	DH (DH + SH inclusive)	SH (SH only)
Model fit, p<0.05	(vs DH) 37.1%	(vs SH) 94.1%	(vs TH) 7%
Model fit, p<0.005	(vs DH) 20.7%	(vs SH) 89.6%	(vs TH) 9.3%
Average AICc	22528.1	22530.9	22691.8
Average log(L)	-11184.7	-11187.1	-11268.6
Average Omega	0.2424	0.2342	0.2443
Average TH rate	0.33	-	-

Table 1. Comparison of summary statistics and parameters across our models.

Conclusions

Our analyses on the SELECTOME (3) indicate a statistically significant contribution to signal by Serine codon island jumping, in part due to the degeneracy of the genetic code. When serine to serine shifts are disallowed we saw a decrease in the TH rate for 88% (data not shown) of the datasets where TH was prefered (over DH ($p < 0.05$)). **Future work aims to further contextualize the biological contributors (6, 7) of MSH in adaptive evolutionary datasets. Especially in species where we are more likely to find expression of error prone polymerases (2). A better understanding of MSH pressures (8, 9) may also delineate false positive inferences of selection acting upon genes (4).** A current implementation of this method is available for HyPhy version (≥2.4) at: https://github.com/veg/hyphy-analyses/tree/master/FitMultiModel

Contact

Alexander G. Lucaci
Alexander.Lucaci@Temple.edu
Department of Biology
Temple University
github.com/aglucaci

Acknowledgements

I am eternally grateful to Sadie Wisotsky and Sergei Pond for their mentorship and stewardship of this project. Additionally, I want to thank the entire Pond lab (ACME), for their ongoing support and camaraderie during the development of this research. Steven Weaver, for making my life easier. Jordan Zehr, for being a world-class mate. Stephen Shank, for his philosophy of science. Brittany Magalis, for sharing her perspective on multiple sequence alignments. Stella Park, for providing my sanity check. Additional thanks to the Temple College of Science and Technology (CST) for providing a research community which fosters this kind of collaborative work and the Institute for Genomics and Evolutionary Medicine (iGEM), for creating a home for cutting-edge thinking on molecular evolution and phylogenetics.

References

1. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. (1994). *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a040152
2. Harris, K., & Nielsen, R. (2014). Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Research*. https://doi.org/10.1101/gr.170696.113
3. Proux, E., Studer, R. A., Moretti, S., & Robinson-Rechavi, M. (2009). Selectome: A database of positive selection. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkn768
4. Venkat, A., Hahn, M. W., & Thornton, J. W. (2018). Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nature Ecology and Evolution*. https://doi.org/10.1038/s41559-018-0584-5
5. Pond, S. K., & Muse, S. V. (2005). Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msi232
6. Löytynoja, A., & Goldman, N. (2017). Short template switch events explain mutation clusters in the human genome. *Genome Research*. https://doi.org/10.1101/gr.214973.116
7. Wang, Q., Pierce-Hoffman, E., Cummings, B. B., Karczewski, K. J., Alföldi, J., Francioli, L. C., … MacArthur, D. G. (2019). Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *BioRxiv*. https://doi.org/10.1101/573378
8. Dunn, K. A., Kenney, T., Gu, H., & Bielawski, J. P. (2019). Improved inference of site-specific positive selection under a generalized parametric codon model when there are multinucleotide mutations and multiple nonsynonymous rates. *BMC Evolutionary Biology*. https://doi.org/10.1186/s12862-018-1326-7
9. Whelan, S., & Goldman, N. (2004). Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics*. https://doi.org/10.1534/genetics.103.023226