The Journal of Open Source Software

scoup: Simulate Codon Sequences with Darwinian Selection Incorporated as an Ornstein-Uhlenbeck Process

Hassan Sadiq [1,2¶] **and Darren P. Martin** [2]

1 Department of Statistics and Actuarial Science, Stellenbosch University, South Africa **2** Institute of Infectious Diseases and Molecular Medicine, Division of Computational Biology, Department of Integrative Biomedical Sciences, University of Cape Town, South Africa ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 20 June 2025
Published: unpublished

Summary

Genetic analyses of natural selection within and between populations have increasingly developed along separate paths. The two important genres of evolutionary biology (i.e. phylogenetics and population genetics) borne from the split can only benefit from research that seeks to bridge the gap. Simulation algorithms that combine fundamental concepts from both genres are important to achieve such unifying objective. We introduce `scoup`, a codon sequence simulator that is implemented in R and hosted on the Bioconductor platform. There is hardly any other simulator dedicated to genetic sequence generation for natural selection analyses on the platform. Concepts from the Halpern-Bruno mutation-selection model and the Ornstein-Uhlenbeck (OU) evolutionary algorithm were creatively fused such that the end-product is a novelty with respect to computational genetic simulation. Users are able to seamlessly adjust the model parameters to mimic complex evolutionary procedures that may have been otherwise infeasible. For example, it is possible to explicitly interrogate the concepts of static and changing fitness landscapes with regards to Darwinian natural selection in the context of codon sequences from multiple populations.

Statement of need

Statistical inference of the extent to which Darwinian natural selection has impacted genetic data commands a healthy portion of the phylogenetic literature (Jacques et al., 2023). Validation of these largely codon-based models relies heavily on simulated data. Given the ever increasing diversity of natural selection inference models that exist (Kosakovsky Pond et al., 2020; Yang, 2007), there is a need for more sophisticated simulators to match the expanding model complexities.

Bioconductor (Gentleman et al., 2004) is a leading platform where peer-reviewed bioinformatic software useful for biological data analyses are hosted. A search of the entries on the platform, in Version 3.19 on 29 October 2024, with keywords including, `codon`, `mutation`, `selection`, `simulate`, and `simulation` returned a total of 72 unique packages out of the 2300 available. None of the retrieved entries was dedicated to codon data simulation for natural selection analyses. Thus, `scoup` is designed on the basis of the mutation-selection (MutSel) framework (Halpern & Bruno, 1998) as an overdue contribution to the void. Software and/or packages for simulating genetic sequences are also rare in the scientific literature (Gearty et al., 2024).

The Journal of Open Source Software

Algorithm

scoup is further unique for at least three reasons. First, it incorporates Darwinian natural selection into the MutSel model in terms of variability of selection coefficients, an extension of an idea from Spielman & Wilke (2015). Second, it directly utilises the concept of fitness landscapes. Third, fitness landscape updates can be executed in either a deterministic or a stochastic format. The stochastic updates are implemented in terms of the more biologically amenable, Ornstein-Uhlenbeck (OU) process (Bartoszek et al., 2017; Uhlenbeck & Ornstein, 1930). A crude summary of how substitution events are executed in scoup is presented in Figure 1.

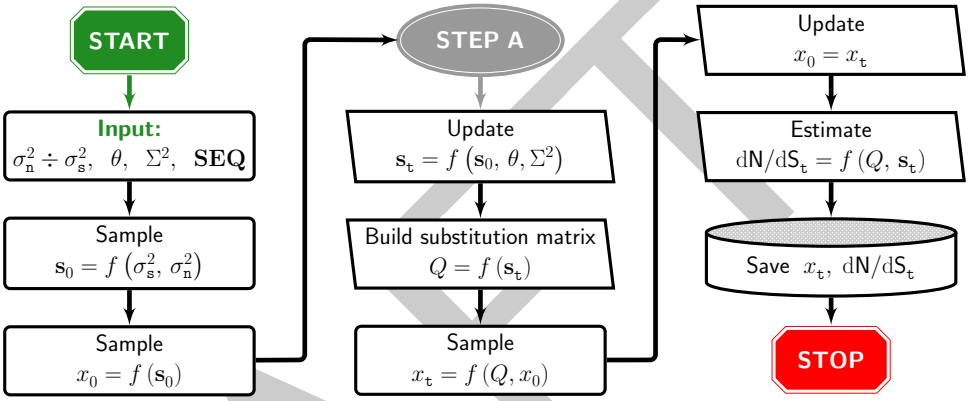

Figure 1: Summarised scoup algorithm. After each substitution event, the process returns to *STEP A*, until the input tree length ($\tau \in \mathbf{SEQ}$) is exhausted. σ_n^2 = variance of amino acid selection coefficients. σ_s^2 = variance of synonymous codon selection coefficients. Σ^2 = OU asymptotic variance. θ = OU mean reversion rate. $\mathbf{SEQ}$ = sequence information. $x_\star$ = codon. $\mathbf{s}_\star$ = codon selection coefficient vector.

Implementation

We simulated (see sample code in Figure 2) 20 independent sequence alignments made up of 1000 codon sites and 8 extant taxa for each of the parameter combinations presented in Figure 3. The phylogeny used was balanced and the length of its branches were 0.10 each. The stochastic OU framework was implemented and other function inputs were left at their default values (Figure 2). Estimates of $\mathrm{d}N/\mathrm{d}S$ were obtained following Spielman & Wilke (2015) and were averaged over all selection coefficient updates at each site and across the alignment. Inferences of ω were obtained with `CODEML` in `PAML` (Yang, 2007).

```
adaptEnrty <- ouInput() # Line01
modelEntry <- hbInput() # Line02
sqEntry <- seqDetails() # Line03
seqData <- alignsim(adaptEnrty, sqEntry, modelEntry)
```

Figure 2: An example R code for simulating a codon sequence alignment with scoup. Default values were left unchanged. `Line01`: OU adaptation parameters where, $\mu = 0$, $\Sigma^2 = 0.01$ and $\theta = 0.01$. `Line02`: evolution model input where, $\mathbf{s} \sim \text{Gamma}(1, \sigma_n^{-1})$, $\sigma_n^2 = 10^{-5}$, $\sigma_s^2 = 10^{-5}$ and effective population size, $N_e = 1000$. `Line03`: sequence information where, site count is 250, extant taxa count is 64 and branch length is 0.1.

The Journal of Open Source Software

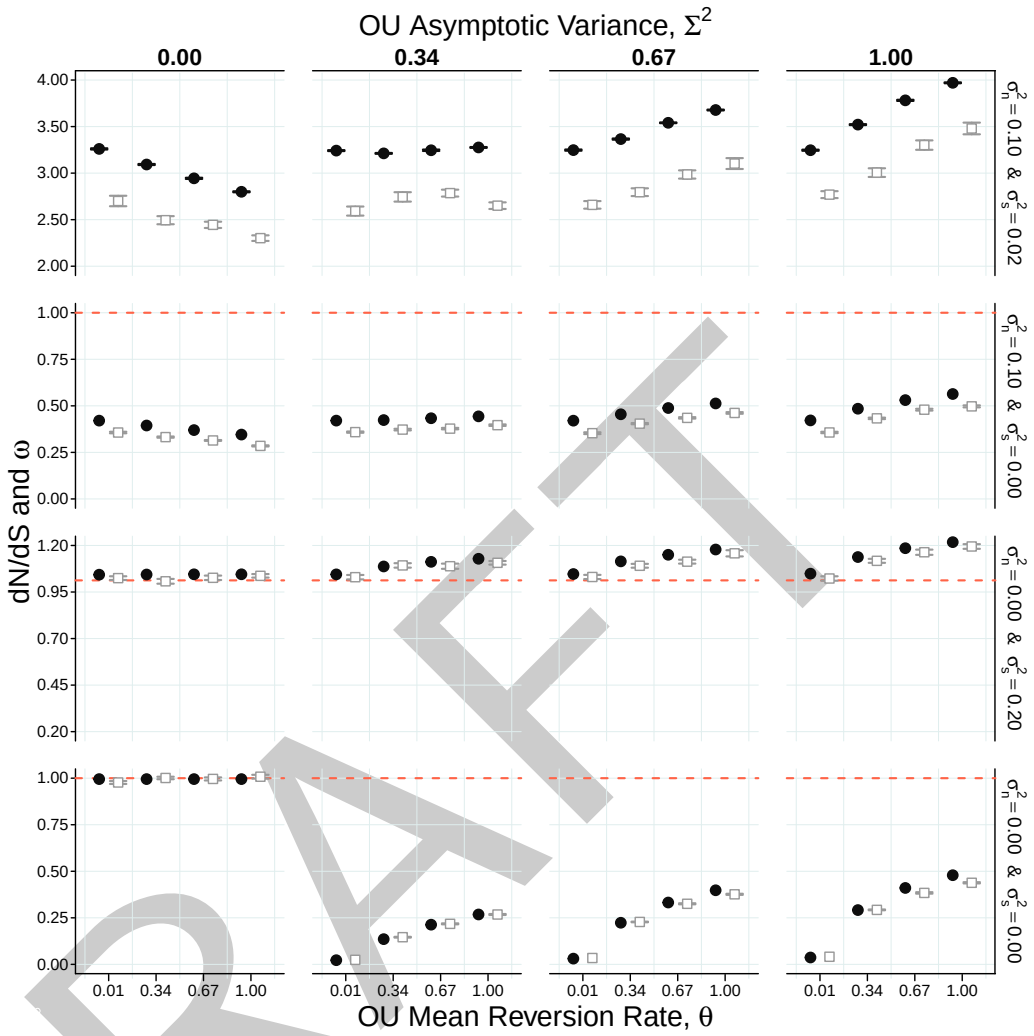

Figure 3: Demonstration of the accuracy of outputs from scoup in terms of the likelihood ω and the analytical dN/dS measures of natural selection. The estimates of the selection measures were obtained homogeneously from each alignment generated for every combination of the stochastic landscape (Σ^2 and θ) and the Darwinian selection (σ_n^2 and σ_s^2) parameters. The filled circles represent the average $\mathrm{d}N/\mathrm{d}S$ estimates while the empty squares represent the average ω estimates, across 20 independent codon sequence alignments. The widths of the arrows correspond to twice the standard errors. The dashed lines highlight point of neutral selection effect.

Estimates of ω and $\mathrm{d}N/\mathrm{d}S$ summarised in Figure 3 strongly agree, except for the case of $(\sigma_n^2, \sigma_s^2) = (0.10, 0.02)$. The suppressed ω estimates, that is most pronounced for $\sigma_n^2, \sigma_s^2 > 0$, is likely a consequence of the well-documented conservative property of homogeneous ω inference techniques (see for example, Nielsen & Yang (1998)). Regardless, a correlation coefficient of approximately 0.9971 was obtained when the ω and $\mathrm{d}N/\mathrm{d}S$ averages were compared. The standard errors ranged between $[0.0000,\ 0.0077)$ and $(0.0004,\ 0.0627)$ for the $\mathrm{d}N/\mathrm{d}S$ and ω estimates respectively. These measures confirm that the outputs from scoup are accurate.

Conclusions

We present scoup, a R package that allows for simulation of codon sequences in a way that is capable of recapitulating the evolutionary processes of biological systems more realistically

than most existing simulators. Our framework creatively incorporates the Ornstein-Uhlenbeck process into the mutation-selection evolutionary model. This attribute could potentially unlock exciting research avenues that will improve existing knowledge about the complex interactions of different, potentially interacting, molecular evolutionary processes. In another unique contribution to the literature, the magnitude of the Darwinian selection affect on the simulated sequences was controlled with the ratio of the variances of selection coefficients. Given the summaries in Figure 3, we state the following hypothesis with respect to natural selection inference from multi-population genetic sequences. With ω, it is difficult to fully distinguish between compensatory and adaptive diversifying selection occurring on static and changing landscapes, respectively. To establish this hypothesis, at least numerically, scoup should be an invaluable resource.

Code availability

scoup is published for free public use under the GPL-2 license. It is available for download from the Bioconductor platform, along with detailed documentation and tutorial files.

Whitepaper

A scoup whitepaper is available on the bioRxiv preprint server.

Acknowledgements

We thank Ben Murrell for suggesting modelling varying selection coefficients with an OU process. Computations were performed using the HPC1 facility at Stellenbosch University, South Africa.

References

Bartoszek, K., Glémin, S., Kaj, I., & Lascoux, M. (2017). Using the Ornstein–Uhlenbeck Process to Model the Evolution of Interacting Populations. *Journal of Theoretical Biology*, *429*, 35–45. https://doi.org/10.1016/j.jtbi.2017.06.011

Gearty, W., O'Meara, B., Berv, J., Ballen, G. A., Ferreira, D., Lapp, H., Schmitz, L., Smith, M. R., Upham, N. S., & Nations, J. A. (2024). *CRAN Task View: Phylogenetics*. https://CRAN.R-project.org/view=Phylogenetics

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., … Zhang, J. (2004). Bioconductor: Open Software Development for Computational Biology and Bioinformatics. *Genome Biology*, *5*(10), R80. https://doi.org/10.1186/gb-2004-5-10-r80

Halpern, A. L., & Bruno, W. J. (1998). Evolutionary Distances for Protein-Coding Sequences: Modelling Site-Specific Residue Frequencies. *Molecular Biology and Evolution*, *15*(7), 910–917. https://doi.org/10.1093/oxfordjournals.molbev.a025995

Jacques, F., Bolivar, P., Pietras, K., & Hammarlund, E. U. (2023). Roadmap to the Study of Gene and Protein Phylogeny and Evolution – A Practical Guide. *PLoS One*, *18*(2), e0279597. https://doi.org/10.1371/journal.pone.0279597

Kosakovsky Pond, S. L., Poon, A. F. Y., Velazquez, R., Weaver, S., Hepler, N. L., Murrell, B., Shank, S. D., Magalis, B. R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman, S. J., Frost, S. D. W., & Muse, S. V. (2020). HyPhy 2.5 – A Customizable Platform

for Evolutionary Hypothesis Testing Using Phylogenies. *Molecular Biology and Evolution*, *37*(1), 295–299. https://doi.org/10.1093/molbev/msz197

Nielsen, R., & Yang, Z. (1998). Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics*, *148*(3), 929–936. https://doi.org/10.1093/genetics/148.3.929

Spielman, S. J., & Wilke, C. O. (2015). The Relationship between dN/dS and Scaled Selection Coefficients. *Molecular Biology and Evolution*, *32*(4), 1097–1108. https://doi.org/10.1093/molbev/msv003

Uhlenbeck, G. E., & Ornstein, L. S. (1930). On the Theory of the Brownian Motion. *Physical Review*, *36*, 823–841. https://doi.org/10.1103/PhysRev.36.823

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088