

# <sup>1</sup> scoup: Simulate Codon Sequences with Darwinian Selection Incorporated as an Ornstein-Uhlenbeck Process

<sup>4</sup> Hassan Sadiq  <sup>1,2</sup>¶ and Darren P. Martin  <sup>2</sup>

<sup>5</sup> Department of Statistics and Actuarial Science, Stellenbosch University, South Africa <sup>2</sup> Institute of  
<sup>6</sup> Infectious Diseases and Molecular Medicine, Division of Computational Biology, Department of  
<sup>7</sup> Integrative Biomedical Sciences, University of Cape Town, South Africa ¶ Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Editor: Julia Romanowska 

## Reviewers:

- [@clauswilke](#)
- [@ogoeli](#)

Submitted: 20 June 2025

Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## <sup>8</sup> Summary

<sup>9</sup> 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.

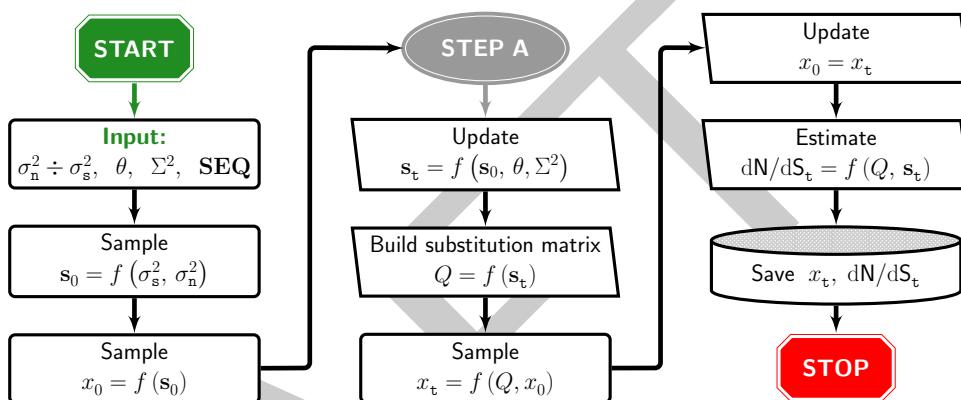
## <sup>23</sup> Statement of need

<sup>24</sup> 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](#)).  
<sup>25</sup> 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.

<sup>30</sup> 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.  
<sup>31</sup> 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](#)).

## Algorithm

<sup>38</sup> scoup is further unique for at least three reasons. First, it incorporates Darwinian natural  
<sup>40</sup> selection into the MutSel model in terms of variability of selection coefficients, an extension  
<sup>41</sup> of an idea from Spielman & Wilke (2015). Second, it directly utilises the concept of fitness  
<sup>42</sup> landscapes. Third, fitness landscape updates can be executed in either a deterministic or a  
<sup>43</sup> stochastic format. The stochastic updates are implemented in terms of the more biologically  
<sup>44</sup> amenable, Ornstein-Uhlenbeck (OU) process (Bartoszek et al., 2017; Uhlenbeck & Ornstein,  
<sup>45</sup> 1930). A crude summary of how substitution events are executed in scoup is presented in  
<sup>46</sup> Figure 1.



**Figure 1: Summarised scoup algorithm.** After each substitution event, the process returns to *STEP A*, until the input tree length ( $\tau \in \text{SEQ}$ ) is exhausted.  $\sigma_n^2$  = variance of amino acid selection coefficients.  $\sigma_s^2$  = variance of synonymous codon selection coefficients.  $\Sigma^2$  = OU asymptotic variance.  $\theta$  = OU mean reversion rate. **SEQ** = sequence information.  $x_*$  = codon.  $s_*$  = codon selection coefficient vector.

<sup>47</sup> We highlight two important design choices from Figure 1. First, we assume that a static fitness  
<sup>48</sup> landscape is obtained from a single set of parameters ( $\xi$ ) needed to sample a 20-element  
<sup>49</sup> numerical vector of amino acid selection coefficients (that is,  $s_0$  in Figure 1). The coefficients  
<sup>50</sup> are subsequently used as inputs of the corresponding MutSel model. This ensured that a  
<sup>51</sup> seascape setting is then defined as a function of multiple sets of parameters ( $\xi_1, \xi_2, \dots, \xi_k$ ,  
<sup>52</sup> where  $k \leq$  extant taxa size). Second, the coefficient update ( $s_t$ ) step is done after every  
<sup>53</sup> substitution event. In addition, the Ornstein-Uhlenbeck update process is discretised. In other  
<sup>54</sup> words, the OU jump sizes are fixed and pre-specified as an input to the simulation functions.

## Implementation

<sup>56</sup> scoup may be installed directly from Bioconductor using the following R code in Figure 2.

```

if(!requireNamespace("BiocManager", quietly=TRUE))
  install.packages("BiocManager")
BiocManager::install("scoup")
  
```

**Figure 2: R installation code for scoup.** Allows the most recently published version of the package to be installed from Bioconductor. Development version of the package may be installed by adding `BiocManager::install(version="devel")` before the final line.

<sup>57</sup> A sample code for executing a simulation run with scoup is presented in Figure 3. The code

58 executes a stochastic OU framework on a balanced phylogeny with 64 extant taxa.

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

**Figure 3: 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,  $s \sim \text{Gamma}(1, \sigma_n^{-1})$ ,  $\sigma_n^2 = 10^{-5}$ ,  $\sigma_s^2 = 10^{-5}$  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.

## 59 Conclusions

60 We present **scoup**, a R package that allows for simulation of codon sequences in a way that  
 61 is capable of recapitulating the evolutionary processes of biological systems more realistically  
 62 than most existing simulators. Our framework creatively incorporates the Ornstein-Uhlenbeck  
 63 process into the mutation-selection evolutionary model. This attribute could potentially unlock  
 64 exciting research avenues that will improve existing knowledge about the complex interactions  
 65 of different, potentially interacting, molecular evolutionary processes. In another unique  
 66 contribution to the literature, the magnitude of the Darwinian selection affect on the simulated  
 67 sequences was controlled with the ratio of the variances of selection coefficients.

## 68 Code availability

69 **scoup** is published for free public use under the GPL-2 license. It is available for download  
 70 from the [Bioconductor platform](#), along with detailed documentation and tutorial files. Some  
 71 additional sample code are accessible in tests/ and vignettes/ folders of the package.

## 72 Acknowledgements

73 We thank Ben Murrell for suggesting modelling varying selection coefficients with an OU  
 74 process. Computations were performed using the [HPC1](#) facility at Stellenbosch University, South  
 75 Africa.

## 76 References

- 77 Bartoszek, K., Glémin, S., Kaj, I., & Lascoux, M. (2017). Using the Ornstein–Uhlenbeck  
 78 Process to Model the Evolution of Interacting Populations. *Journal of Theoretical Biology*,  
 79 429, 35–45. <https://doi.org/10.1016/j.jtbi.2017.06.011>
- 80 Gearty, W., O'Meara, B., Berv, J., Ballen, G. A., Ferreira, D., Lapp, H., Schmitz, L., Smith,  
 81 M. R., Upham, N. S., & Nations, J. A. (2024). CRAN Task View: Phylogenetics.  
<https://CRAN.R-project.org/view=Phylogenetics>
- 83 Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B.,  
 84 Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R.,  
 85 Leisch, F., Li, C., Maechler, M., Rossini, A. J., ... Zhang, J. (2004). Bioconductor: Open  
 86 Software Development for Computational Biology and Bioinformatics. *Genome Biology*,  
 87 5(10), R80. <https://doi.org/10.1186/gb-2004-5-10-r80>

- 88 Halpern, A. L., & Bruno, W. J. (1998). Evolutionary Distances for Protein-Coding Sequences:  
89 Modelling Site-Specific Residue Frequencies. *Molecular Biology and Evolution*, 15(7),  
90 910–917. <https://doi.org/10.1093/oxfordjournals.molbev.a025995>
- 91 Jacques, F., Bolivar, P., Pietras, K., & Hammarlund, E. U. (2023). Roadmap to the Study  
92 of Gene and Protein Phylogeny and Evolution – A Practical Guide. *PLoS One*, 18(2),  
93 e0279597. <https://doi.org/10.1371/journal.pone.0279597>
- 94 Kosakovsky Pond, S. L., Poon, A. F. Y., Velazquez, R., Weaver, S., Hepler, N. L., Murrell,  
95 B., Shank, S. D., Magalis, B. R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman,  
96 S. J., Frost, S. D. W., & Muse, S. V. (2020). HyPhy 2.5 – A Customizable Platform  
97 for Evolutionary Hypothesis Testing Using Phylogenies. *Molecular Biology and Evolution*,  
98 37(1), 295–299. <https://doi.org/10.1093/molbev/msz197>
- 99 Spielman, S. J., & Wilke, C. O. (2015). The Relationship between dN/dS and Scaled Selection  
100 Coefficients. *Molecular Biology and Evolution*, 32(4), 1097–1108. <https://doi.org/10.1093/molbev/msv003>
- 102 Uhlenbeck, G. E., & Ornstein, L. S. (1930). On the Theory of the Brownian Motion. *Physical  
103 Review*, 36, 823–841. <https://doi.org/10.1103/PhysRev.36.823>
- 104 Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology  
105 and Evolution*, 24(8), 1586–1591. <https://doi.org/10.1093/molbev/msm088>

DRAFT