The Journal of Open Source Software

TREvoSim v3: An individual based simulation for generating trees and character data

Russell J. Garwood [1,2¶], **Alan R. T. Spencer** [2,3], **Charles T. M. Bates** [1], **Leah M. Callender-Crowe** [4], **Frances S. Dunn** [5], **Thomas J. D. Halliday** [6], **Joseph N. Keating** [7], **Nicolás Mongiardino Koch** [8], **Luke A. Parry** [9], **Robert S. Sansom** [1], **Thomas J. Smith** [9], **Mark D. Sutton** [3], **and Thomas Vanteghem** [10]

1 Department of Earth and Environmental Sciences, University of Manchester, Manchester, M13 9PL, UK **2** Natural History Museum, London, SW7 5BD, UK **3** Department of Earth Science and Engineering, Imperial College, London, SW7 2AZ, UK **4** School of Biological Sciences, University of Reading, Reading, UK **5** Oxford University Museum of Natural History, University of Oxford, Oxford, OX1 3PW, UK **6** School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham B15 2TT, UK **7** School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, UK **8** Scripps Institution of Oceanography, UC San Diego, La Jolla, CA 92122, USA **9** Department of Earth Sciences, University of Oxford, Oxford, OX1 3AN, UK **10** Erasmus Mundus Joint Master Degree PANGEA, Université de Lille, France ¶ Corresponding author

DOI: 10.21105/joss.06722

Software

- Review
- Repository
- Archive

Editor: Frederick Boehm

Reviewers:

- @atcribb
- @ms609

Submitted: 25 March 2024
Published: 10 September 2024

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

Summary

Simulations provide valuable insights into the patterns and processes of evolution, and the performance of analytical methods used to investigate empirical data (Barido-Sottani et al., 2020; Dolson & Ofria, 2021; Wright & Hillis, 2014). Here we present TREvoSim v3.0.0: an agent-/individual-based model written in C++, in which digital organisms evolve, creating phylogenetic character data and trees. Trees inferred from empirical data always carry uncertainty, but TREvoSim can create a known tree alongside associated character data, allowing – for example – assessment of phylogenetic inference methods (Keating et al., 2020; Mongiardino Koch et al., 2021, 2023). The v3.0.0 release adds a range of features to allow greater flexibility in simulating phylogenetic trees and character data (more logging options, finer control over character character and simulation parameters), and to facilitate the study of broader evolutionary topics (e.g. ecosystem engineering, adaptive landscapes, selection).

Background

A range of digital platforms to study evolution and ecology exist, with varied approaches and levels of abstraction (Dolson & Ofria, 2021). TREvoSim is a sister-package to the spatially explicit eco-evolutionary simulation REvoSim (Furness et al., 2023; Garwood et al., 2019). v1.0.0 was developed to investigate the accuracy and precision of phylogenetic inference methods (Keating et al., 2020). After further development, TREvoSim v2.0.0 was used to investigate the impact fossils have on phylogenetic inference and evolutionary timescales (Mongiardino Koch et al., 2021, 2023). In brief, TREvoSim is a non spatially-explicit model in which organisms – which consist of a genome of binary characters – compete within a structure called the playing-field to echo natural selection (Figure 1). Their chance of replication is dictated by a fitness algorithm that assesses organismal fit against a series of random numbers (masks, constituting an environment). On replication, organisms have a chance of mutation, and descendents overwrite a current member of the playing field. The simulation has a lineage-based species concept, and at the end of a simulation the software can output trees

and characters (species genomes), as well as logging the simulation state as the model runs.

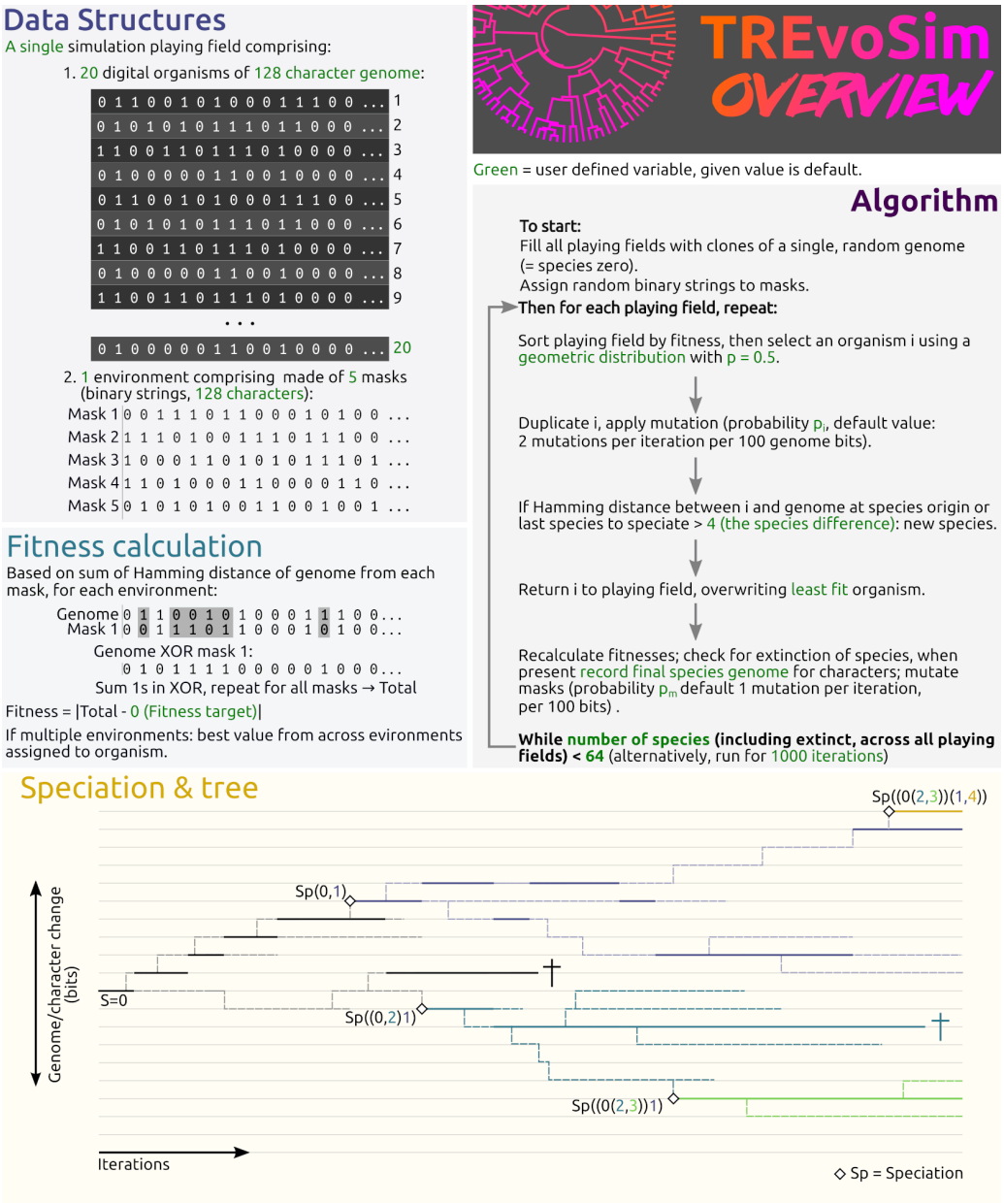

Figure 1: Figure 1 - A simplified overview of TREvoSim. Green text represents a user-defined variable and the given value is default. The figure is split into data structures, the fitness calculation, the algorithm, and the tree. In the tree, character change is represented by change on the Y axis, time on the X. Any lineage is likely to comprise multiple individuals: modal genomes are shown in solid lines, and non-modal via dashed lines. By default, genomes for each species are recorded on their extinction. A full description of the algorithm is available in the TREvoSim documentation.

New features

TREvoSim v3.0.0 includes a suite of new features that allow the investigation of an expanded range of evolutionary processes. These new features are briefly introduced (in alphabetical order) below, and fully described in the TREvoSim documentation.

Model enhancements

Ecosystem engineering

A new ecosystem engineering system allows the impact of organism-environment feedback to be investigated (Jones et al., 1994). When the ecosystem engineering option is enabled (it is disabled in default settings), a species is assigned ecosystem engineering status halfway through a run, and passes this status to descendents. When this occurs, the genome of that organism is either used to overwrite an environmental mask, or added to the environment as an additional mask. Overwriting a mask reduces the Hamming distance between engineers and masks, and – assuming a low fitness target – thus directly improves their fitness relative to non-engineers. In contrast, adding a mask changes the nature of the fitness landscape for all organisms, but with a weaker direct benefit to ecosystem engineers. Ecosystem engineering can occur just once ('one-shot' ecosystem engineering) or can be repeated after the first application ('persistent'). When a mask is added in the first application, it is overwritten in subsequent applications when ecosystem engineers are persistent. A new facility to log the ecosystem-engineering status of individuals is provided.

Expanding playing field

This new option alters competition such that it occurs between species rather than between individuals. When enabled (disabled by default), this is achieved by allowing only one individual from each species to be present in the playing field at any time. As such, the playing field grows to accommodate new species, which appear following the standard speciation rules. On duplication of an individual, juveniles overwrite the previous member of their species in the playing field. Otherwise, the playing field operates as normal, i.e. it is ordered by fitness and duplication of an individual selected using a geometric distribution links fitness to fecundity. Enabling the expanding playing field removes intra-species competition from the simulation: this can be used to investigate the impact that intra- vs. inter-species competition has on evolutionary processes during a run, and on the phylogenetic outcomes of a simulation.

Match peaks

When there are multiple environments for any given playing field, by default masks are seeded with random numbers, and may thus have different peak fitness values. This new option (disabled by default) instead seeds each playing field with environments that have the same peak fitness. This is achieved by doing a site-wise randomisation of the sequence of zeros and ones across masks, between environments. For example, with three masks in a given environment, the first site may be 1,0,0 for masks 1,2 and 3 respectively – when this option is enabled, the pattern from site one may be moved to site seven between the first and second environment, and this is repeated for all sites. This operation ensures that the best achievable fitness at the start of a simulation will remain the same, but will be achieved by a different genome across environments. In v3.0.0, as the simulation progresses and mutations to the masks occur, matching peaks are no longer guaranteed (it is possible this will change in future releases). Additionally, the simulation uses a heuristic algorithm to generate an initial seed organism that has the same fitness in each environment (in $>99\%$ of cases) when this option is selected. In general, this option allows finer control of the fitness landscape, and its impact on evolution to be investigated.

No selection

Another new addition is a no selection mode – when this is enabled (disabled by default), organisms for replication are chosen from the playing field at random, rather than using fitness to determine replication probability. When this option is enabled, the simulation functions under drift, allowing study of e.g. neutral vs. selective regimes.

Playing field mixing

By default TREvoSim playing fields are independent data structures, and organisms in one playing field do not compete with those in others during a simulation. This new option allows configurable mixing of organisms between playing fields, which can be asymmetrical if desired. Playing field mixing can facilitate the study of, for example, the dynamics of invasive species or biotic interchanges.

Stochastic layer

Provides a layer of abstraction between an organism's genome and the bits used for the fitness calculation (this option is disabled by default). It achieves this using many-to-one mapping (i.e. using a sequence of multiple bits to define the value of a single bit), the map being defined by the user. For example, the sequence of bits 0110 in a genome might map to a 1 bit in the genome calculation, whilst are 1100 may map to 0. As such, any individual bit does not necessarily have an impact on the fitness of an organism as a whole, allowing e.g. neutral mutations and less strongly adaptationist dynamics.

Perturbations

When enabled (disabled by default), this implements a limited period of increased rates of environmental change that occurs halfway through a run (when half the requested species have evolved, or at half the requested iteration number). This is intended for study of scenarios where evolutionary dynamics are driven by variations in the rate of environmental change(Condamine et al., 2013). There is an option to also increase mixing between playing fields during a perturbation. This system can provide insights into the impact of disturbance and rapid environmental change, for example, on evolution in a non-spatially explicit setting.

Software modifications

Character limits

New options allow finer control of TREvoSim functions that employ genome characters. Characters in TREvoSim are used in several portions of the algorithm – they form the basis of calculating fitness of organisms, and are also employed in the identification of species. In previous versions of TREvoSim, all characters were used for both functions, through a user-defined total character number. From v3, a separate limit on the character count used for species selection and/or the fitness calculation can be applied (in the default settings, all are the same). When either – or both – differ from the total character number, only a subset of characters (those from zero to the limit) are included in the defined operations, and others can evolve independent of these processes (i.e., in the absence of selective forces, akin to more neutral drift-like processes).

Default simulation parameters

New default values for simulation parameters are introduced with this release. Outputs created using these default variables have been compared against twelve empirical, total evidence analyses (following the approach of, and using empirical data sourced from, Mongiardino Koch et al., 2021) for three measures of tree shape and homoplasy. The empirical data, analysis script, and resulting graphs have been placed within the source code repository, and the results are presented in the documentation.

Running log

In previous versions, TREvoSim only allowed outputs at the end of a run. From v3, an additional logging system has been added, called the running log, that allows the user to create a customised file recording many aspects of the simulation state – for example, the tree

and the character data – during a run. The user can define all log outputs using the logging options (which are outlined in the documentation), and can also dictate the frequency with which the running log is written.

Simulation modes

Simulation-termination can now be configured to occur after either a user-defined number of iterations, or when a user-defined number of species have evolved; only the latter was previously available. This provides increased flexibility in experimental design.

Codebase Tests

TREvoSim v3.0.0 introduces a test suite covering all aspects of the simulation mechanics. These can be accessed by the user, through the graphical user interface, by developers through an integrated development environment such as Qt Creator, and by both through running the test binary, as outlined in the documentation.

Statement of need

TREvoSim employs a selection-driven, agent-based approach: it incorporates key elements of biological evolution (selection, reproduction, heritability and mutation). The (true) phylogenetic trees and character data are an emergent property of a TREvoSim simulation, and as such the software is particularly well suited to simulation studies that can be analysed through phylogenetic trees and character data matrices. These include, for example: the impact of missing data on phylogenetic inference; the impact of rates of environmental change on character evolution; and the nature of evolution under different fitness landscapes. This complements other simulation approaches where, typically, phylogenetic trees or character data are simulated using stochastic tools such as birth-death models (e.g. Guillerme, 2024) or data based on random numbers (e.g. Puttick et al., 2019) that do not incorporate, for example, selection acting on individuals. The data TREvoSim generates are different in a number of ways to those created using stochastic models (Keating et al., 2020), and are also likely to violate the assumptions of models commonly used in phylogenetic inference. Incorporating a level of model misspecification resembling that expected from empirical datasets is desirable in simulation studies that assess the efficacy of inference methods. Given the complexity of morphological evolution, the subsequent impact of character coding, and our current understanding of the patterns present in empirical character data, it is challenging to demonstrate, beyond the inclusion of empirically grounded concepts in its generation, the naturalism of TREvoSim data. The default settings have been chosen to reflect a number of features of empirical data matrices and trees to try and minimise the mismatch between TREvoSim and real world data, however, there are a broad range of potential alternative means of quantifying outputs. Which of these is most appropriate is likely to depend on the area of study and specific question at hand, and as such, TREvoSim provides granular control over the simulation parameters, allowing users to generate data that best serve their needs. TREvoSim is intended as a versatile platform that might be used to study a broad range of topics.

Current associated projects

In addition to the published studies, cited previously, future directions include:

- TJS, FSD, LAP, RJG - The Macroevolutionary consequences of ecosystem engineering.
- CTMB, TJS, LAP, RJG, FSD - Fitness landscapes and disparity.
- TJDH, RJG - Mixing and perturbations.
- RSS, ARTS, RJG, JNK - Impact of character number and correlated characters in a phylogenetic context.

- NMK, LAP, RJG - Phylogenetic method development.
- JNK, RJG - Impact of evolutionary mode on phylogenetic character data.
- MDS, RJG - Fossilisation and phylogenetic inference.

Author contributions

RJG developed and coded TREvoSim, with support on testing and releasing from ARTS. RSS, LMC-C, MDS, and JNK contributed ideas during the initial phases and continued development of the software, and JNK proposed the stochastic layer. Peak matching and associated features were developed in collaboration with CTMB, FSD, LAP and TJS, and ecosystem engineering was developed in collaboration with LAP, FSD and TJS. Mixing and perturbations were proposed by TJDH. Development of many of the features released in v2.0.0 was conducted in collaboration with, and empirical benchmarking was led by, NMK. The expanding playing field was developed with and analysed by TV.

Availability

TREvoSim v3.0.0 source code and binaries are freely available from Zenodo and GitHub. Newer releases of the TREvoSim software will be available on GitHub. Full documentation is available from ReadTheDocs.

Acknowledgements

RJG and RSS were supported by the NERC award NE/T000813/1 and development began during the BBSRC grant BB/N015827/1 awarded to RSS and RJG. RJG is supported by the Alexander von Humboldt Foundation; RJG, LAP, and RSS by Leverhulme Trust Research Project Grant (RPG-2023-234); FSD by NERC fellowship NE/W00786X/1; and LAP by NERC fellowship NE/W007878/1. We are grateful to the reviewers of this contribution, Martin R. Smith and Alison T. Cribb, for suggestions that significantly improved the quality of this publication and the TREvoSim documentation, and to Martin R. Smith for additional auditing of our code.

References

Barido-Sottani, J., Saupe, E. E., Smiley, T. M., Soul, L. C., Wright, A. M., & Warnock, R. C. M. (2020). Seven rules for simulations in paleobiology. *Paleobiology*, *46*(4), 435–444. https://doi.org/10.1017/pab.2020.30

Condamine, F. L., Rolland, J., & Morlon, H. (2013). Macroevolutionary perspectives to environmental change. *Ecology Letters*, *16*, 72–85. https://doi.org/10.1111/ele.12062

Dolson, E., & Ofria, C. (2021). Digital evolution for ecology research: A review. *Frontiers in Ecology and Evolution*, *9*. https://doi.org/10.3389/fevo.2021.750779

Furness, E. N., Garwood, R. J., & Sutton, M. D. (2023). REvoSim v3: A fast evolutionary simulation tool with ecological processes. *Journal of Open Source Software*, *8*(89), 5284. https://doi.org/10.21105/joss.05284

Garwood, R. J., Spencer, A. R. T., & Sutton, M. D. (2019). REvoSim: Organism-level simulation of macro and microevolution. *Palaeontology*, *62*(3), 339–355. https://doi.org/10.1111/pala.12420

Guillerme, T. (2024). Treats: A modular r package for simulating trees and traits. *Methods in Ecology and Evolution*, *15*(4), 647–656. https://doi.org/10.1111/2041-210X.14306

Jones, C. G., Lawton, J. H., & Shachak, M. (1994). Organisms as ecosystem engineers. *Oikos*, *69*(3), 373–386. https://doi.org/10.2307/3545850

Keating, J. N., Sansom, R. S., Sutton, M. D., Knight, C. G., & Garwood, R. J. (2020). Morphological phylogenetics evaluated using novel evolutionary simulations. *Systematic Biology*, *69*(5), 897–912. https://doi.org/10.1093/sysbio/syaa012

Mongiardino Koch, N., Garwood, R. J., & Parry, L. A. (2021). Fossils improve phylogenetic analyses of morphological characters. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1950), 20210044. https://doi.org/10.1098/rspb.2021.0044

Mongiardino Koch, N., Garwood, R. J., & Parry, L. A. (2023). Inaccurate fossil placement does not compromise tip-dated divergence times. *Palaeontology*, *66*(6), e12680. https://doi.org/10.1111/pala.12680

Puttick, M. N., O'Reilly, J. E., Pisani, D., & Donoghue, P. C. J. (2019). Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. *Palaeontology*, *62*(1), 1–17. https://doi.org/10.1111/pala.12388

Wright, A. M., & Hillis, D. M. (2014). Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE*, *9*(10), e109210. https://doi.org/10.1371/journal.pone.0109210