Research Ideas and Outcomes 10: e124988
doi: 10.3897/rio.10.e124988

Forum Paper

Prototype Biodiversity Digital Twin: Phylogenetic Diversity

Vladimir Mikryukov[‡], Kessy Abarenkov[§], Thomas S. Jeppesen[|], Dmitry Schigel[|], Tobias Guldberg Frøslev[|]

‡ Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia
§ University of Tartu Natural History Museum, Tartu, Estonia
| Global Biodiversity Information Facility - Secretariat, Copenhagen Ø, Denmark

Corresponding author: Tobias Guldberg Frøslev (tfroeslev@gbif.org)

Reviewed v 1

Academic editor: Sharif Islam

Received: 08 Apr 2024 | Accepted: 10 Jun 2024 | Published: 11 Jun 2024

Citation: Mikryukov V, Abarenkov K, Jeppesen TS, Schigel D, Frøslev T (2024) Prototype Biodiversity Digital Twin: Phylogenetic Diversity. Research Ideas and Outcomes 10: e124988. https://doi.org/10.3897/rio.10.e124988

Abstract

Phylogenetic diversity (PD) represents a fundamental measure of biodiversity, encapsulating the extent of evolutionary history within species groups. This measure, pivotal for understanding biodiversity's full dimension, has gained recognition by major environmental and scientific organisations, including the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Unlike traditional taxonomic richness, PD offers a comprehensive, evolutionary perspective on biodiversity, essential for conservation planning and biodiversity management. This manuscript describes the development of a BioDT (Biodiversity Digital Twin) prototype, aimed at facilitating the calculation and visualisation of biodiversity metrics from global, dynamic data sources. By utilising the PhyloNext pipeline and integrating with global phylogenetic and species occurrence databases like the Open Tree of Life (OToL) and the Global Biodiversity Information Facility (GBIF), the prototype aims to significantly reduce computation time and enhance user interaction. This enables dynamic visualisation and potentially hypothesis testing, making it a valuable tool for researchers, monitoring initiatives, policy-makers and the public. The prototype's development focuses on improving the PhyloNext pipeline's scalability and creating a more intuitive user interface, expanding its utility for conservation efforts and biodiversity exploration. Our work illustrates the potential impact of the BioDT

prototype in supporting diverse user groups in visualising and exploring PD, thus contributing to more informed decision-making in conservation and biodiversity management.

Keywords

biodiversity metrics, evolution, conservation, PhyloNext, digital twin, phylogenetic diversity

Introduction

Phylogenetic diversity (PD) quantifies the extent of evolutionary history encompassed by a group of species, highlighting a crucial dimension of biodiversity. "Acknowledged by the Intergovernmental Science Policy Platform on Biodiversity and Ecosystem Services (IPBES, www.ipbes.net/), the Earth's evolutionary legacy is considered a vital component of biodiversity, safeguarding possibilities for future generations. The tree of life serves as a repository of possible advantages for humans and through the preservation of PD, we protect the diversity of traits (essentially, the wide array of evolutionary characteristics found within a group of species) and secure future opportunities for human benefit". (IUCN SSC Phylogenetic Diversity Task Force, https://www.pdtf.org/). Biodiversity is most commonly quantified through taxonomic richness. For example, it is common to describe how diverse a genus or a geographic area is by counting the number of species within them. On the other hand, PD, a metric that takes into account the branch lengths in a phylogenetic tree, provides an evolutionary perspective of biodiversity that cannot be estimated using species richness alone. PD (expected loss of phylogenetic diversity) was one of the proposed indicators for the Kunming-Montreal Global Biodiversity Framework (https://www.cbd.int/doc/decisions/cop-15/cop-15-dec-05-en.pdf). When combined with geographical constraints, a model utilising PD metrics can effectively compare and qualify competing areas as most relevant for the designation or expansion of protected natural areas.

The Global Biodiversity Information Facility (GBIF) is an intergovernmental network and biodiversity data infrastructure, that currently mediates almost 3 billion species occurrences. The Open Tree of Life (OToL) provides a comprehensive, accessible and continuously updated synthesised phylogenetic tree amongst all known species. The PhyloNext pipeline integrates these two pivotal research data infrastructures, making them more accessible to non-experts, by generating PD metrics. The PhyloNext pipeline (Mikryukov et al. 2024) generates PD metrics using the Biodiverse programme (Laffan et al. 2010). PhyloNext uses Docker containers allowing for relatively easy local installation, but is also possible to launch in a cloud environment. PhyloNext is operated by commands in a terminal. A demo graphical interface is hosted by GBIF.org. The tool can be used to explore PD across different geographic and/or taxonomic groups in relation to policy-making, prioritisation of conservation efforts etc. However, even relatively simple queries – for example, Felidae (cats) in South Africa – usually require several hours of computation and the tool cannot presently be used easily as an interactive exploratory visualisation tool

in near-real-time. With a significant improvement of computation time and an improved - and perhaps simplified - user interface, the tool would become invaluable for visualising/ testing conservation strategies or simply exploring PD. Furthermore, by incorporating additional information – for example, shapefiles with geographical strata or classes – the tool may be extended to address specific hypotheses/questions, such as whether areas designated as nature reserves, based on species richness, also support high PD. The work on this Phylogenetic Diversity Digital Twin will focus on up-scaling the existing PhyloNext pipeline and creating a more intuitive user interface that targets the most common and relevant use cases. This will create a responsive user experience allowing the tool to be used to, for example, identifying localities to survey and data from such surveys would feed into the next updating of the model. We envision a future version to have hypothesis-testing modules, allowing for comparative evaluation of alternative poposals for, for example, designating protected nature areas.

Objectives

The objective of this prototype is to leverage the Biodiverse programme as implemented in the PhyloNext pipeline to develop a tool that facilitates calculation and visualisation of PD metrics and other biodiversity metrics from large, standardised and dynamic global data sources with a time expenditure that potentially allows for dynamic visualisation adequate for interactive exploration and refinement of input procedures (and potentially parameters) to achieve interactive fine-tuning of output. A central developmental focus will be on upscaling the calculations to achieve near real-time outputs of metrics. Another focus will be to use (either directly or as inspiration) the existing demo interface to develop a user interface that is intuitive to use for experts and non-experts alike and devise a dynamic and interactive visualisation module for refinement and optimisation of input parameters. The development will consider possible enhancements like hypothesis testing, based on shape files. We envision that this prototype will serve users across various domains from researchers, monitoring initiatives, policy-makers and interested citizens with interest in natural history and biodiversity.

Workflow

The existing GBIF demo interface (https://phylonext.gbif.org/) is a graphical user interface developed for the PhyloNext pipeline. It allows the user to provide settings in six sections of a web-based form. The first section, **Name and description**, allows the user to set a name and description for the particular pipeline/model to be run with selected setting. In the **Phylogeny** section, the user can choose to upload/provide a custom phylogenetic tree in Newick format or to use a few select pre-defined trees, based on the open tree of life. Information on the format taxon labels is also required. In the **Taxonomic filters** section, the user defines whether the model should be restricted to certain taxa (at any of the classic taxonomic levels: phylum, class, order, family genus). In **Spatial and temporal filters**, a range of years can be defined to which occurrences should be restricted, as well as spatial constraints in the form of a hand-drawn polygon, country name or an uploaded

polygon in the GeoPackage format (https://www.geopackage.org/). In the section **GBIF Occurrence filtering and aggregation**, a number of filters can be engaged to filter the GBIF occurrence data to exclude likely flawed data (e.g. occurrences with known suspicious coordinates of museums, country capitals, country centroids etc.) and types of data (e.g. material samples), likely spatial outliers identified through density-based clustering. Finally, the section **Biodiverse settings** allows the user to define the parameters of the Biodiverse programme that calculates the metrics from the filtered data. After starting the pipeline with the defined settings, a significant amount of time is needed before the user has a visual output.

The envisioned workflow of the final BioDT prototype will be based on improvement of the demo interface with a focus on user friendliness and simplification and aiming for a more interactive, dynamic process, where the user can fine-tune parameters, based on the initial output. A hypothesis-testing module would likely be a combination of a graphical part and text, both for input and output.

The conceptual schema of the proposed workflow is shown in Fig. 1.

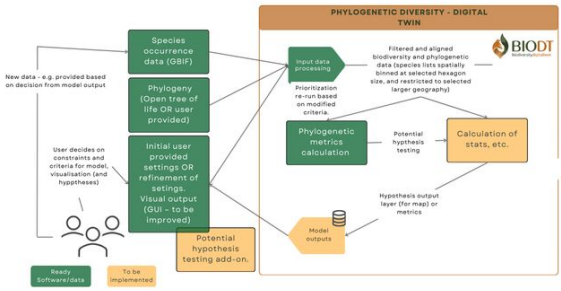

Figure 1. doi

Schematic workflow.

Data

The PhyloNext pipeline is already accessible and in practice able to use all global occurrences records from GBIF. Currently, these amount to almost 3 billion occurrences of taxa across the tree of life. Theoretically, all these taxa would also be in the Open Tree of Life phylogeny. However, in reality, there are several groups – especially of bacteria, fungi and micro-eukaryotes – where the taxonomy is in flux and there are many species that do

not carry formal binomial names. Several of these groups are best defined by molecular species concepts and there have emerged several systems for making such dark taxa operational by providing them with persistent, unique identifiers. As these identifiers slowly find their way into both global molecular phylogenies and biodiversity databases, we will be able to obtain Phylogenetic Biodiversity metrics that vastly surpass what is currently available both in accuracy and taxonomic coverage and with much less bias. The data sources are described in Table 1.

Table 1.

Data sources.

Data source	Data type	Notes
Global phylogeny from Open tree of life. (https://tree.opentreeoflife.org/)	Phylogenetic tree (Newick format representing graph-theoretical trees with edge lengths using parentheses and commas)	The synthetic phylogenetic tree from the OToL, constructed using all the contributing trees, is available for download in Newick format. PhyloNext accesses OToL through web APIs and retrieves a taxon-specific tree, optionally filtering out taxa without phylogenetic support. The latest released synthetic tree (v. 14.9) includes 2,392,578 tips (≈ taxa/species).
Global taxon occurrences (2+ billion) from the Global Biodiversity Information Facility (https://www.gbif.org/)	Species occurrence data in GBIF is standardised according to the Darwin Core Standard (https://dwc.tdwg.org/) maintained by the Darwin Core Maintenance Interest Group of the Biodiversity Information Standards (TDWG, https://www.tdwg.org/)	The PhyloNext pipeline is compatible with any data dump from GBIF. GBIF provides full monthly data dumps of all species occurrences (www.gbif.org/occurrence-snapshots). Currently, GBIF mediates 2,950,644,031 occurrence records (observations/detections of taxa).

Model

The final prototype biodiversity twin will, in most aspects, reuse the analytical flow and tools developed in the PhyloNext pipeline.

Key features of the model:.

- Input data specifications: As described in the workflow, the model takes filtered occurrence data from GBIF and phylogenetic data from the Open Tree of Life.
- Phylogenetic tree preparation: The workflow supports pre-constructed phylogenetic trees, as well as retrieving synthetic trees from the Open Tree of Life. This step includes matching species names (from the tips of phylogenetic) with GBIF species keys.
- Spatial Binning: The workflow uses a discrete global grid system – for example, H3 by Uber (https://h3geo.org/) for the spatial binning of species occurrences.

- Diversity and endemism estimation: Using the Biodiverse programme (Laffan et al. 2010) for each grid cell of the study area, the workflow calculates an array of diversity metrics.

PhyloNext uses Nextflow (https://www.nextflow.io) to run tasks in a workflow. Docker containers (https://www.docker.com/) facilitates easy installation.

PhyloNext and its dependencies are fully documented and described in the GitHub repository (https://phylonext.github.io/). The source code for the demo interface GUI (https://phylonext.gbif.org/) is openly available in two GitHub repositories (backend: https://github.com/gbif/phylonext-ws; frontend: https://github.com/gbif/phylonext-ui). All the phylogenetic diversity estimates and other biodiversity indices and metrics are estimated by the Biodiverse programme that can calculate more that 300 metrics. Biodiverse code and installation are documented in a GitHub repository (https://shawnlaffan.github.io/biodiverse/).

A schematic illustration of the dataflow in PhyloNext can be see in Fig. 2. Image from Mikryukov et al. (2024).

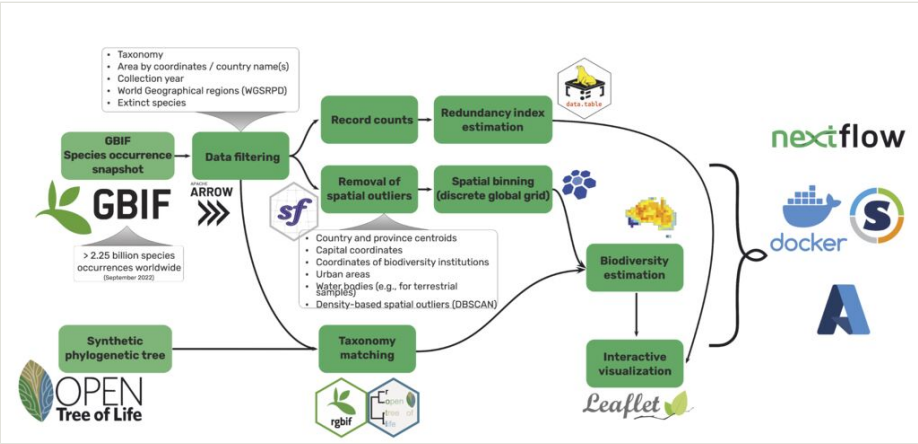

Figure 2. doi

PhyloNext data flow and model. Illustration from Mikryukov et al. (2024).

FAIRness

The occurrence data used for the modelling is FAIR in the sense that it is all publicly available as standardised data in the GBIF index under open licences and findable and accessible via web interfaces (https://www.gbif.org) and APIs (https://techdocs.gbif.org/en/openapi/). The demo interface offers the functionality to generate a citable DOI for each PhyloNext pipeline execution and to create a sharable link to the results. Additionally, GBIF has implemented a mechanism for generating a unique DOI for a derived dataset (https://www.gbif.org/derived-dataset/about), facilitating tracking and proper accreditation of all

individual datasets that contributed to the underlying occurrence data. Similarly, the data from OToL is licensed openly and accessible with APIs (https://github.com/OpenTreeOfLife/germinator/wiki/Open-Tree-of-Life-Web-APIs). The source code for the PhyloNext pipeline and the demo GUI are also fully open and accessible on GitHub (see above). The code to be developed for the BioDT prototype will also be fully open, ensuring its alignment with FAIR principles. A hosting at GBIF.org will make the DT fully open and available and will be provided with a CC-BY-NC licence. The outputs of the model will be able to download using widely accepted standards for biodiversity and spatial data.

Performance

Due to the design of PhyloNext, which utilises containerisation technologies (Docker and Singularity/Apptainer) to encapsulate all software dependencies, the pipeline was successfully installed on the petascale supercomputer LUMI. The initial tests have also been conducted with success. The next steps involve exploring various approaches for scaling up. The PhyloNext pipeline is highly adaptable for HPC and cloud environments. For example, it allows for the configuration of specific resource requirements (CPUs and RAM) for each process independently, allowing the pipeline to launch these tasks as independent jobs through SLURM workload manager. Alternatively, a fixed amount of resources (e.g. a single computational node) can be allocated exclusively for the pipeline's operation, facilitating the optimisation for specific tasks of datasets. For storage, the pipeline is also capable of utilising S3-compatible object storage which can significantly enhance performance by offering scalable, high-speed access to data. Additionally, a resource usage profiler is included, which allows us to monitor and optimise the resources required for the analysis. Furthermore, the Biodiverse programme, which calculates a wide array of diversity metrics, incorporates essential optimisations (e.g. caching and re-using computationally expensive calculations) that can significantly enhance the speed of analysis.

Interface and outputs

As describe above, the demo interface is a web-based user interface with a panel where users can select the geography and taxonomy of interest (e.g. mammals of South Africa), choose a phylogenetic tree and configure model settings (the size of spatial bins, the number of randomisation iterations etc.). Results are being shown on a map. Currently, some implicit knowledge about the procedure is expected from the user: for example, different types of taxon labels and what bin sizes, number of randomisations and what the various optional filtering terms in Darwin Core (the TDWG biodiversity data standards used by GBIF) mean and which values they can have. Additionally, it is easy for a user without prior knowledge to select parameters that conflict (e.g. using a tree that has a taxon focus other than the taxonomic filter for the occurrence data). The user interface of the Digital Twin prototype to be developed is intended to be a more user-friendly version of that GUI, for example, with more guidance, protection against conflicting values and fixed vocabularies where it makes sense. A number of pre-defined models (specific settings of

the model) is also planned to allow an approach for users to start with examples, that intuitively make sense.

Integration and sustainability

If the final prototype runs smoothly and is user-friendly as planned, one possibility is hosting it at the Global Biodiversity Information Facility (GBIF.org), similar to the current demo interface, but with a formal release and comprehensive user guidance.

Application and impact

If a fast and user-friendly version of PhyloNext, equipped with an intuitive graphical user interface, is successfully developed, the Digital Twin prototype may become a valuable tool and analytical hub for many years. The primary data sources – occurrence data from GBIF.org – is constantly growing and supported by a stable infrastructure. Currently, the model by default uses the synthetic phylogenetic tree from the Open Tree of Life project, a resource that is also continuously expanding and improving. Thus, the estimates produced by the model will automatically improve over time.

A fast and interactive tool for visualising phylogenetic diversity (and other associated metrics) may serve numerous applications across various user groups as mentioned above. Researchers may use it to quickly visualise and explore taxonomic groups or geographic areas of interest as a tool to formulate new hypotheses. Monitoring initiatives may use the tool to visualise the impact of their work and identify areas of future attention or sampling. Policy-makers will be able to examine the potential impact of competing proposals for nature conservation.

By integrating modules for hypothesis testing and other advanced functionalities, the scope of potential applications could expand even further. Examples include comparative studies across ecosystems (e.g. comparisons across different ecosystems or biomes, which can provide information for conservation priorities and strategies at both the European level and globally), evaluation of the impact of different agricultural practices on biodiversity, invasive species management (e.g, identifying potential hotspots for invasive species spread and assessing effectiveness of management strategies), ranking of potential areas for the expansion of nature reserves and various analyses across time series or other stratifications (e.g. data types).

Acknowledgements

We thank Joseph T. Miller (GBIF) for his visionary conception and support of the development of the PhyloNext pipeline, Shawn Laffan (UNSW Sydney) for the development of Biodiverse and his invaluable assistance with its integration into the workflow, Emily Jane McTavish (UC Merced) for the expert assistance in fetching phylogenetic trees from the OToL, Tim Robertson (GBIF) and Matthew Blissett (GBIF) for

their support of the GBIF infastructure and for managing access to the data, John Waller (GBIF) for his work in exploring the species occurrence filtering workflow and Tuomas Rossi (CSC) for technical assistance with the LUMI supercomputer. The deveopment of PhyloNext was supported by a grant “PD (Phylogenetic Diversity) in the Cloud” to GBIF Supplemental funds from the GEO-Microsoft Planetary Computer Programme. This study has received funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101057437 (BioDT project, https://doi.org/10.3030/101057437). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Conflicts of interest

The authors have declared that no competing interests exist.
Disclaimer: This article is (co-)authored by any of the Editors-in-Chief, Managing Editors or their deputies in this journal.

References

- Laffan SW, Lubarsky E, Rosauer DF (2010) Biodiverse, a tool for the spatial analysis of biological and related diversity. Ecography 33 (4): 643-647. https://doi.org/10.1111/j.1600-0587.2010.06237.x
- Mikryukov V, Abarenkov K, Laffan S, Robertson T, McTavish EJ, Jeppesen TS, Waller J, Blissett M, Kõljalg U, Miller JT (2024) PhyloNext: a pipeline for phylogenetic diversity analysis of GBIF-mediated data. URL: https://phylonext.github.io/