Prototype Biodiversity Digital Twin: prioritisation of DNA metabarcoding sampling locations

Tobias Guldberg Frøslev[‡], Robin James Boyd[§], Dmitry Schigel[‡]

‡ Global Biodiversity Information Facility - Secretariat, Copenhagen Ø, Denmark
§ UK Centre for Ecology & Hydrology, Wallingford, United Kingdom

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

Academic editor: Volker Grimm

Abstract

Advancements in environmental DNA (eDNA) metabarcoding have revolutionised our capacity to assess biodiversity, especially for cryptic or less-studied organisms, such as fungi, bacteria and micro-invertebrates. Despite its cost-effectiveness, the spatial selection for sampling sites remains a critical challenge due to the considerable time and resources required for processing and analysing eDNA samples. This study introduces a Biodiversity Digital Twin Prototype, aimed at optimising the selection and prioritisation of eDNA sampling locations. Leveraging available eDNA data and integrating user-defined criteria, this digital twin facilitates informed decision-making in selecting future sampling sites. Through the development of an associated data formatting tool, we also facilitate the accessibility and utility of DNA metabarcoding data for broader conservation efforts. This prototype will serve multiple end-users, from researchers and monitoring initiatives to commercial enterprises, by providing an intuitive interface for interactive exploration and prioritisation, based on estimated complementarity of future samples. The prototype offers a scalable approach to biodiversity sampling. Ultimately, this tool aims to refine our understanding of global biodiversity patterns and support targeted conservation strategies through efficient eDNA sampling.

Keywords

DNA, metabarcoding, optimisation, modelling, monitoring

Introduction

Environmental DNA metabarcoding and other DNA-based methods are highly efficient in targeting organism groups that normally receive little attention, for example, fungi, bacteria, archaea, protists, nematodes and micro-invertebrates (e.g. Arribas et al. (2021), Kirse et al. (2021)). Methods are simple and cost-effective, also at scales where

traditional sampling regimes would be too costly in terms of money, time and labour (Taberlet et al. 2012). However, to be able to include such data in global biodiversity conservation efforts, it is necessary to achieve at wider global sampling and understand diversity in cryptic environments better and to target areas and localities that are likely to complement current knowledge in the best possible way.

Although methods like DNA metabarcoding are cost effective compared to alternatives for detecting cryptic diversity, each sample is still costly and time-consuming and there is still a relatively large time span from sampling to a curated list of detected organisms. Compared to, for example, the larger organisms (vascular plants and vertebrates) in a particular habitat, it is not possible to quickly gauge whether a potential sampling area or locality is suitable, diverse, unique or representative in terms of the composition of cryptic species. Thus, it is important to develop more informed strategies for selecting sampling areas and localities when the target is cryptic organisms and the aim is to expand current knowledge with, for example, likely complementary sampling.

Existing approaches to identifying locations at which to collect new biodiversity data tend to aim for an improvement in the performance of some model. Examples include targeting sites that are expected to have substantial statistical leverage over model parameters (Callaghan et al. 2019), sites where the variance model predictions are high (DECIDE tool) and sites that would alter the predictions of a species distribution model the most (Flint et al. 2024). Similar model-based approaches have been proposed in other disciplines (Andersson et al. 2023). The limitation of these methods is that they are conditional on the chosen model, so essentially assume that it will be used for all future analyses.

In this Biodiveristy Digital Twin Prototype, we aim to identify how a digital twin can be used to identify and prioritise areas and locations for further sampling guided from currently available data from similar sampling methods, in combination with user-defined criteria/constraints. To facilitate the uptake of more DNA-derived biodiversity data of relevance to the prototype, we developed a data formatting tool to reshape DNA metabarcoding data to GBIF indexable Darwin Core Archives.

Objectives

The objective of this Digital Twin prototype is to develop a tool to guide/help select future sampling sites, through a relatively easy-to-use interface that allows the user to set constraints and select target areas and organism groups and through a few interactive steps ending with a series of suggested areas or localities to conduct future sampling. We envision that the prototype will be of use for researchers, monitoring initiatives and commercial companies. Research projects, for example, often target cryptic organisms with eDNA metabarcoding and try to use estimated complementarity of future samples as the main or secondary criterion for selecting sampling locations/sites. Monitoring initiatives may use the tool to help designate localities/areas for further sampling according to selected criteria like spatial coverage, estimated complementarity,

representation of habitat types etc. Commercial companies are moving into the area of offering services to large enterprises like mining companies and offer to do monitoring of representative areas for monitoring, but we also see companies offering biodiversity monitoring systems financed by the novel development of digital tokens linked to monitoring biodiversity with environmental DNA samples (e.g. SimplexDNA franklins). All these need a tool to help identify relevant sampling locations. The objective for the associated data formatting/publishing tool is to facilitate uptake from the research community to increase the amount of relevant data available for the prototype. The audience here is data holders, producers and persons involved in biodiversity data publishing in research infrastructures like GBIF, OBIS and ALA.

Workflow

The envisioned workflow will be that the user defined an areas of interest (e.g. a country, a hand-drawn polygon, a shapefile), optionally combined with some strata of interest, based on default available maps of land-cover/use, ecoregions, habitat-types etc., at the scale of interest: for example, "calcareous grassland and heath" if the scope is national/ regional or "tropical humid forest and tropical dry forest" if the scope is global). The user also defines the taxonomic scope (e.g. bacteria, agaricomycetes, mammals, everything based on eDNA) and decides if the data should be binned and at what bin size and which criteria should be used as a basis for prioritisation (and their relative weights). Criteria can be simple metrics per sample/bin, like spatial distance to known samples and (estimated) species richness, but can also be more complex like number of known samples per statum, sample heterogeneity within strata, sample coverage within strata etc. The model then calculates the relevant biodiversity metrics of the relevant samples (within geographic and taxonomic scope) and then calculates predicted priority/ rank of the spatial bins and presents a map with the spatial bin coloured by their rank. An expected example scenario could be that spatial bins (locations) with a high physical distance to known samples, estimated to have a high complementary species pool within the stratum of interest are ranked high (e.g. some remote locations with raised bogs in Denmark). The user will then refine the ranking by modifying criteria and deselecting spatial bins on the map, until a satisfactory and realistic set of localities are left. When the identified areas are then sampled and data are published to GBIF, the model then has access to more data and will be improved.

For the initial version of the prototype, we plan to use one (or a few) homogeneous dataset known to the authors, in combination with a land-use/cover map relevant to the initial tests. These suggestions are listed in Table 1. However, the vision is to be able to source all relevant GBIF data, construct a dataset shape from those and be able to use any relevant map with spatial strata.

The envisioned workflow is illustrated in Fig. 1.

Data

Table 1

The digital twin prototype will ultimately be able to use all global eDNA-metabarcoding derived occurrences records from GBIF. Currently, these amount to some 10-20 million (or more) occurrences. Many of these occurrences come from MGnify, but are currently not formatted for easy reuse in this prototype as the sequence itself is lacking from the data. However, it is estimated that the coming years will see a large influx of suitable data readily usable for this prototype. Eventually, the model will pool data of relevance into spatial bins and use these bins as the geographical unit. With the more sparse initial test data, spatial binning is not needed as eDNA metabarcoding data are sampling event data, where each sample results in a long list of OTUs or detected taxa and, thus, suitable for community compositional biodiversity metric calculations (like richness, dissimilarity scores like Bray-Curtis etc.) when combined with other samples into contingency tables. The intial version of the prototype will use a dataset of soil eDNA from Denmark (see Table 1) or another comparable dataset with internally standardised samples. Some advanced uses will employ stratification of the samples into, for example, habitat types inferred from spatial metadata of the sample. For the initial test, the model will use a land-cover map of Denmark, but, ultimately, any other source of spatial statification should be possible to use. Ultimately, it should also be possible to include more explanatory variables, for example, in the shape of topographic information, climate and weather data, to enhance the efficiency of the predictions.

Model

The user can set basic constraints for future samples, and then use existing data to predict where further samples are best placed. Information from new samples will eventually feed into the model and provide an updated map of priority areas.

The first version will likely use pre-formatted OTU table data. Later versions will use occurrence data from GBIF filtered to include only relevant data.

Input data processing

All existing (eDNA-derived) occurrence data from GBIF is accessed either via APIs or one of the full monthly data dumps of all species occurrences (or the user provides one dataset that should for the basis of the prediction/prioritisation. This approach is used for the initial phase where one dataset is used, see above). Optional maps or other sources of spatial stratification are also made available to the model. The occurrence data (OTUs) are then filtered by the selected geographical and taxonomic scope, grouped at the selected spatial bin size – for example, using the Uber's H3 system (hexagonal hierarchical spatial index) – and each spatial bin is assigned to a geographic stratum, if relevant.

Community metrics calculation

The dataset now consists of a number of spatial bins – some with corresponding pooled data and some without data. For each bin (with data), metrics are calculated (e.g. species richness, uniqueness and some beta diversity metrics). Metrics per stratum (e.g. land-use category) are also calculated: for example, coverage, average community dissimilarity (beta dispersion), average uniqueness.

Prioritisation

All the calculated metrics per bin (with data) are now combined into a calculated rank or score, based on the chosen criteria and their weighting. All (data) empty spatial bins are then assigned a predicted score, based on the known data. In the case where the model has very few parameters – for example, only two spatial strata are provided (forest/not forest), all the "blank" bins receive a score equal to the average score of "known" bins within that stratum. More elaborate modelling of the predicted scores could be envisioned if the model is allowed to draw on other spatial information that can be used as explanatory vaiables (e.g. soil type, climate and weather data). These predicted scores (of "blank" spatial bins) are then weighted by their distance to known samples to make use of the overall default priority to fill spatial gaps when sampling.

User Interface/Output

The output is presented as a map with the spatial bins ranked (coloured) by their predicted rank (priority). The user may ideally interactively modify the prioritisation criteria and their weights. It should also be possible to deselect spatial bins/areas (exclude them from prioritisation) to trigger recalculation of metrics and scores. The visualisation should be an interactive (zoomable etc.) map with spatial bins coloured according to score/rank and a panel with some overall statistics. The aim here is to provide the user with the best locations for optimising global/regional/regional knowledge (gap identification) for further biodiversity sampling given some constraints in study design.

Validation

Though the validation method is under development, possible options are to do a leave-one-out validation - for example, run the model with some set constraints and criteria and see how well the predicted scores (and/or biodiversity metrics) in spatial bins (or samples) align with the actual measurements. Filling spatial gaps of knowledge is likely one of the preferred criteria in any realistic user situation and is a subjective metric enforced by the user (and not dictated by the data themselves). However, all the data-derived metrics should be possible to evaluate with this approach.

FAIRness

We aim for documenting the entire provenance chain from the input data, to use processing steps and modelling to the generated output. Occurrence data used for the

modelling is FAIR in the sense that it is all openly available in GBIF.org, but the exact dataset used will also be defined by a persistent identifier for full reproducibility. Topographic data and other data will be referenced by persistent identifiers. All data used in the models will be available publicly and free to use and share (with appropriate credit). Outputs from DT and the modelling tools to produce the outputs will also be publicly available.

The code for the associated data formatting tool is fully openly available in GitHub (frontend and backend).

Performance

Our initial plan is to base the model on community dissimilarity metrics like Bray-Curtis dissimilarities etc. Although such metrics are more easily calculated than, for example, joint species distributions, the required computational power is significant as the size of species/site matrices (sparse contingency tables by default) grow to enormous size with increased species and sites in the input data. This is further amplified as the model here targets hyperdiverse communities of microbial species. Even with access to a supercomputer, we will likely have to explore alternative ways to handle these large datasets.

Interface and outputs

Though the graphical user interface is not yet developed, we aim to build a graphical user interface with a combination of a panel where the user can define focus areas, constraints etc. from dropdown lists and forms and a panel with a map, showing the areas of interest.

The user interface for the prototype of the associated data formatting/publishing tool can be seen here.

Integration and sustainability

If a completed prototype operates smoothly using data sourced from GBIF, it could potentially be hosted on GBIF.org.

Application and impact

If a user tool is successfully developed that allows users to obtain informed and prioritised suggestions for future potential sampling areas or localities for eDNA samples, this Digital Twin prototype may potentially have a broad user base. The underlying primary data source – the DNA metabarcoding derived occurrence data from GBIF.org – has only been available in a standardised and interoperable form for a few years, but the

potential growth is enormous due to decreased costs in sequencing, better analytical tools and easier options for data sharing. Thus, the predictions and suggestions of the model will be automatically improved vastly by increased data availability. Researchers may use it to quickly help selecting locations for sampling even with varied sampling designs. Monitoring initiatives may use the tool to locate single locations or larger areas of interest, where sampling gaps exist. Commercial companies that offer biodiversity data generation to customers like multinational corporations abiding by regulatory requirements on impact on nature, organisations engaging in impact certificates, companies creating data for ecological assessments etc.

Considering that eDNA is a growing pool of data and we live in a rapidly changing world, we see a potential of future models also being able to include a temporal aspect, to be able to identify areas of great change that may be prioritised for monitoring to detect and document change in biodiveristy.

Acknowledgements

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.

References

- Andersson TR, Bruinsma WP, Markou S, Requeima J, Coca-Castro A, Vaughan A, Turner RE (2023) Environmental sensor placement with convolutional Gaussian neural processes. Environmental Data Science 2: 32. https://doi.org/10.1017/eds.2023.22
- Arribas P, Andújar C, Salces-Castellano C, Castellano A, Emerson BC, Vogler AP (2021) The limited spatial scale of dispersal in soil arthropods revealed with whole-community haplotype-level metabarcoding. Molecular Ecology 30 (1): 48-61. https://doi.org/10.1111/mec.15591
- Callaghan CT, Poore AGB, Major RE, Rowley JJL, Cornwell WK (2019) Optimizing future biodiversity sampling by citizen scientists. Proceedings. Biological sciences 286 (1912): 20191487. https://doi.org/10.1098/rspb.2019.1487
- Flint I, Wu CH, Valavi R, Chen WJ, Lin TE (2024) Maximising the informativeness of new records in spatial sampling design. Methods in Ecology and Evolution 15 (1): 178-190. https://doi.org/10.1111/2041-210X.14260

- Kirse A, Bourlat SJ, Langen K, Fonseca VG (2021) Unearthing the potential of soil eDNA metabarcoding-Towards best practice advice for invertebrate biodiversity assessment. Frontiers in Ecology and Evolution 9: 630560. https://doi.org/10.3389/fevo.2021.630560
- Levin G (2022) Documentation of the data and method for the elaboration of a land use and land cover map for Denmark. Aarhus University. Technical Report No. 252. DCE – Danish Centre for Environment and Energy. URL: https://envs.au.dk/en/research-areas/society-environment-and-resources/land-use-and-gis/basemap/basemap04-geotiff-for-download
- Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular ecology 21 (8): 2045-50. https://doi.org/10.1111/j.1365-294X.2012.05470.x

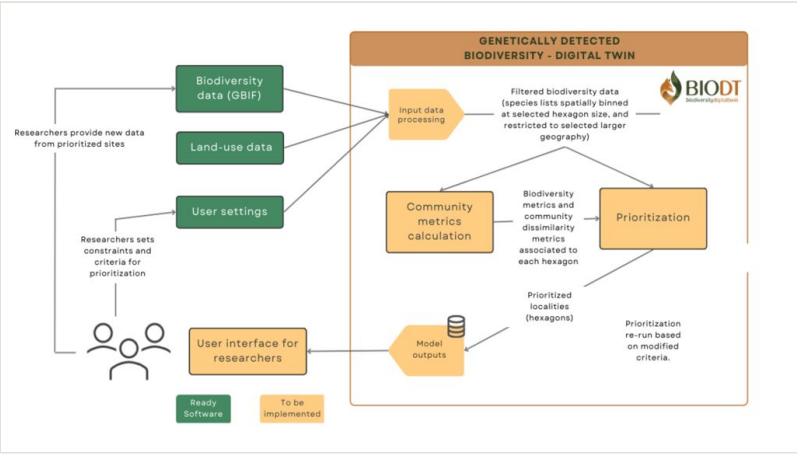

Figure 1.

The DNA metabarcoding Sampling Prioritisation tool in a nutshell.

Table 1.

Data sources.

Data source	Data type	Notes
eDNA fungi DK, example dataset	species occurrence data (OTU table data)	To be used as a pre-formatted dataset as an example. The dataset is also available in GBIF as an occurrence dataset.
GBIF.org	species occurrence data (in this case restricted to DNA-detected data)	Accessible through APIs Some development needed to restrict the data to only eDNA and to pool the data and get it into a shape relevant for the biodiversity calculations.
Land cover map of Denmark (Levin 2022)	Land-use/type data	A published digital map, that needs some processing to become a shape file or raster file suitable for the modelling. Downloadable here Other land-use/type maps may be relevant and needed for later versions to address larger areas. (e.g. a map of global ecoregions).