Intermediate report

Naomi van Es
s1114618

December 2021

Contents

1 Introduction

1.1 ARISE-project

To better understand the biodiversity in the Netherlands, the ARISE project has been established[1]. The University of Amsterdam, Naturalis Biodiversity Center, the University of Twente and the Westerdijk Fungal Biodiversity Institute are working together to create an infrastructure to identify and monitor all multicellular species found in the Netherlands. Data processed with artificial intelligence (e.g., radar, images, radio) and eDNA analysis are used to achieve this.

1.2 Species Identification

Biodiversity is often measured as the different (number) of species found in an area, even though it is also measured by looking at the differences in genes within a species.[2] To ascertain which organisms are in any given area, species detection is necessary. Organisms are put in their respective taxa using a universal nomenclature based on the Linnean hierarchy. To which taxon an organism is assigned can be based on morphological traits, although this method can be error-prone and time-consuming.[3] A different approach is using DNA barcoding, first introduced by Hebert et al. in 2003.[4] Taxa are identified using parts of different genes that are distinct enough to establish which species the sequence is derived from. A barcode is generally around 600 base pairs and needs to possess conserved sequences on both ends of the barcode sequence. The conserved sequences are used for primers and are necessary to make the polymerase chain reaction (PCR) possible. PCR multiplies DNA to prepare it for sequencing, which can be performed on various (high-throughput) sequencing platforms. The sequence of interest is compared to a barcode reference library to conclude which species the sequence belongs to. If it is impossible to classify an organism to species level, it can be assigned to a higher taxonomic class like genus or family.[5]

While DNA barcoding is often used to identify one organism, it can also be used on environmental DNA (eDNA) instead. This process is called metabarcoding. Environmental samples like water or soil contain a multitude of organisms that can be identified. After PCR, using universal or general primers, DNA is sequenced using high-throughput sequencing, and the species in the sample are determined. Metabarcoding makes it possible to assess species composition and to monitor biodiversity without disrupting ecosystems.[6] The difference in the general workflow between DNA barcoding and metabarcoding is shown in figure 1.

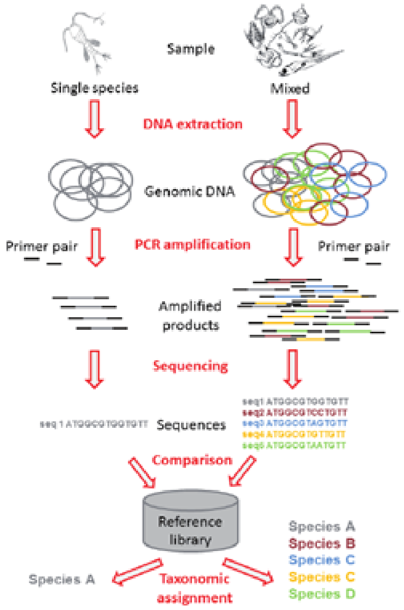

Figure 1: An overview of the workflow using single sample (DNA barcoding) species identification versus a sample with multiple species (metabarcoding) identification.[7]

Ideally, there would be one DNA barcode to identify all species. Unfortunately, this barcode does not exist. However, there are (combinations of) gene markers widely used to separately identify animals, plants, fungi, protists or bacteria. Different DNA sources like nuclear, chloroplast and mitochondrial are used for this. Various combinations of barcodes are suited for different lifeforms. For animals, the locus cytochrome C oxidase I (COI), located on mitochondrial DNA, is widely used as a barcode to identify animal sequences. The COI locus is also referred to as 'the barcode of life'.

The COI barcode is insufficient to identify plants, as they do not have enough SNPs in their mitochondrial genome to identify them to species level. The gene maturase K (matK), located on chloroplast DNA, results in better species identification, especially when combined with other chloroplast genes.[8] For fungi, it is also not recommended to use the COI region. Some fungi contain too many introns in the gene, and there are too few differences between species to make identification possible. Some fungi, like Neocallimastigomycota, do not even have mitochondrial DNA, making COI not suitable to use as a barcode. The loci ITS and RPB1 are used among others to identify fungi.[8][9]

1.3 Custom database

Located in Leiden and existing for over 200 years, Naturalis Biodiversity Center focuses on understanding and conserving biodiversity. Naturalis is a research institute and museum with an extensive collection of over 40 million objects.[10] A student from the University of Applied Sciences Leiden, Bastiaan Anker, has constructed a relational database while serving his internship at Naturalis. This project aimed to periodically complement Naturalis data from the Collectieregistratiesysteem[11] (CRS), using publicly available Barcode of Life Data system[12] (BOLD) reference data. The Entity Relation Diagram (ERD) of the custom relational database made during this project is shown in figure 2.

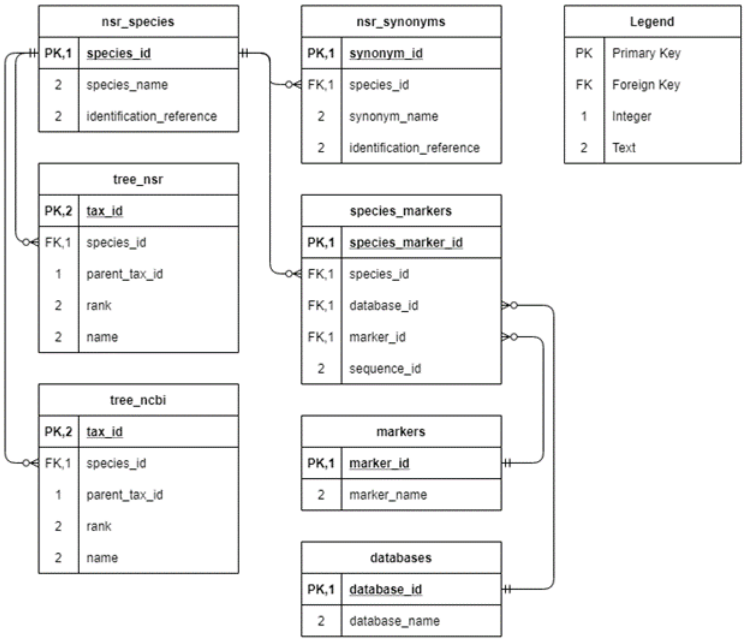

Figure 2: *ERD of the custom database by Anker. The legend shows how primary and foreign keys are presented in the tables. The numbers 1 and 2 show if a data field in the table contains numbers or text. Foreign keys define the relations across tables.[13]*

The custom database consists of seven tables. Data was exported from the Nederlands Soortenregister[14] (NSR) and BOLD. The NSR contains specimen data from organisms native to the Netherlands, whereas the main focus of BOLD is DNA barcoding. The NSR has a checklist including all accepted species found in the Netherlands. BOLD data that included species from the NSR checklist were kept, the rest discarded. Overlap between CRS data and BOLD was filtered. The backbone taxonomy from The National Center for Biotechnology Information[15] (NCBI) and NSR are stored and connected to other tables by a unique species ID. The public and internal specimen records (species_marker) have identifiers to make it possible to retrieve from which databases the records were exported and where to retrieve their metadata.[13]

As mentioned, Naturalis is part of the ARISE project. This project aims to identify and monitor all multicellular species in the Netherlands, so it would be valuable to refine and expand the existing custom database made previously by Anker. Several obstacles are faced when combining reference data. Different sources will lead to inconsistent annotation, for instance, concerning taxonomic assignment. The custom database accommodates NCBI and NSR taxonomy. This needs to be expanded to accommodate the Westerdijk Fungal Biodiversity Institute's backbone taxonomy to make it possible to integrate their data as well. Westerdijk Fungal Biodiversity Institute has a collection of more than 100.000 strains of fungi and bacteria.[16] Because there is no universal sequence identifier, it differs across databases. A new unique identifier is needed to build a BLAST reference database. This will make it possible to link the sequence data to the custom database. BLAST stands for Basic Local Alignment Search Tool, used for searching similarities between nucleotide or protein sequences.[17]

Naturalis has a backlog of internally generated DNA barcodes that need to be indexed. The barcodes are generated using Sanger sequencing and MinION sequencing, a high-throughput sequencing technique with an accuracy of around 56%-88%.[18]. MinION sequences are clustered, and a consensus is created to account for the higher error rate using this sequencing method. Most barcodes from external databases are at species level and sequenced with the Sanger method, with which a single sample can be sequenced to a length of a thousand base pairs, with an accuracy up to 99.999%.[19]

1.4 Research question and aim of this study

This study aims to make a reference FASTA/BLAST database containing DNA barcode sequences, making it possible to determine an environmental sample's taxonomic composition using BLAST.

To access the taxonomy and other relevant metadata, a relational database is needed to link the reference database. A prototype database made by a former bioinformatician intern is to be refined and expanded. The database needs to accommodate the backbone taxonomy of Westerdijk Fungal Biodiversity Institute and additional NSR data. An overview of the BLAST results is to be presented in a user-friendly report in the Galaxy environment of Naturalis. This needs to refer to external databases or the CRS database in case of an internal hit.

Because the reference database will contain existing (Sanger) DNA barcodes as well as newly generated barcodes using MinION sequencing, the central question is if the sequencing technique used will create bias. If there is a bias towards barcodes sequenced with the Sanger method, it could indicate the MinION barcodes are not species-specific.

2 Materials and methods

Relational database:

- Changes in structure relational database (ORMs, extra tables/shifting tables)
- Retrieving data/metadata for BOLD, NSR etc.
- Accommodating backbone WFBI
- Adding barcode sequences
- Extra methods?

Materials:

- B. Ankers prototype database on Github (link)
- Python
- important pkg: psycopg2, SQLalchemy. More pkgs found in requirements.txt Github
- SQL postgress
- BASH/Ubuntu
- Data: BOLD, NSR?, WFBI?, internally generated barcodes

Reference database:

- Constructing reference db
- Retrieve sequences
- Make headers and put into FASTA format

Materials:

- blast v. xx.xx.xx
- bash/python
-

BLAST:

- HTML REPORT

3 Results

4 Discussion

2. Differences in two ERD tables

5 Conclusion

References

[1] Arise-biodiversity.nl. https://www.arise-biodiversity.nl/about. Accessed September 1, 2021.

[2] Rafferty John P. Biodiversity loss - Ecological effects. Encyclopedia Britannica. https://www.britannica.com/science/biodiversity-loss/Ecological-effects 2019. Accessed September 1, 2021.

[3] Cain A. Taxonomy Definition, Examples, Levels, and Classification. Encyclopedia Britannica. https://www.britannica.com/science/taxonomy 2020. Accessed September 6, 2021.

[4] Hebert Paul, Cywinska Alina, Ball Shelley, Dewaard Jeremy. Biological identifications through DNA barcodes. Proc R Soc Lond Ser B Biol Sci *Proceedings. Biological sciences / The Royal Society.* 2003;270:313-21.

[5] Kress W. J., Erickson D. L.. DNA barcodes: Genes, genomics, and bioinformatics *Proceedings of the National Academy of Sciences.* 2008;105:2761–2762.

[6] Ruppert Krista M., Kline Richard J., Rahman Md Saydur. Past, present, and future perspectives of environmental DNA (Edna) metabarcoding: A systematic review in methods, monitoring, and applications of Global Edna *Global Ecology and Conservation.* 2019;17.

[7] Corell Jon, Rodriguez-Ezpeleta Naiara. Tuning of protocols and marker selection to evaluate the diversity of zooplankton using metabarcoding *Revista de Investigación Marina.* 2014;21:19-39.

[8] RS Purty, S Chatterjee. DNA barcoding: An effective technique in molecular taxonomy. Austin Pusblishing Group. https://austinpublishinggroup.com/biotechnology-bioengineering/fulltext/ajbtbe-v3-id1059.php 2016. Accessed September 13, 2021.

[9] Xu Jianping. Fungal DNA barcoding *Genome.* 2016;59:913–932.

[10] 200 jaar Naturalis. Naturalis. https://www.naturalis.nl/200-jaar-naturalis/. Accessed September 6, 2021.

[11] Bioportal. Naturalis https://www.naturalis.nl/en/bioportal. Accessed September 7, 2021.

[12] Bold Systems v4. Boldsystems.org. http://www.boldsystems.org/. Accessed September 7, 2021.

[13] Ankers Bastiaan. *Biodiversity assessment in Dutch freshwater and saltwater areas.* PhD thesis 2021.

[14] Home. Nederlands Soortenregister. https://www.nederlandsesoorten.nl/. Accessed September 13, 2021.

[15] National Center for Biotechnology Information. Ncbi. https://www.ncbi.nlm.nih.gov/. Accessed September 8, 2021.

[16] Collection. Westerdijk Fungal Biodiveristy Institute https://wi.knaw.nl/page/Collection. Accessed September 13, 2021.

[17] Basic Local Alignment Search Tool. Blast. https://blast.ncbi.nlm.nih.gov/Blast.cgi. Accessed September 8, 2021.

[18] Lu Hengyun, Giordano Francesca, Ning Zemin. Oxford Nanopore Minion Sequencing and Genome Assembly *Genomics, Proteomics amp; Bioinformatics.* 2016;14:265–279.

[19] Victoria Wang Xin, Blades Natalie, Ding Jie, Sultana Razvan, Parmigiani Giovanni. Estimation of sequencing error rates in short reads *BMC Bioinformatics.* 2012;13.