Biodiversity Information Science and Standards 3: e35165
doi: 10.3897/biss.3.35165

Conference Abstract

Current situation of DNA Barcoding data in biodiversity and genomics databases and data integration for museomics

Takeru Nakazato [‡]

[‡] Database Center for Life Science, Mishima, Japan

Corresponding author: Takeru Nakazato (nakazato@dbcls.rois.ac.jp)

Received: 04 Apr 2019 | Published: 18 Jun 2019

Citation: Nakazato T (2019) Current situation of DNA Barcoding data in biodiversity and genomics databases and data integration for museomics. Biodiversity Information Science and Standards 3: e35165. https://doi.org/10.3897/biss.3.35165

Abstract

The museomics activity regards museum-preserved specimens as rich resources for DNA studies by extracting and analyzing DNA from these specimens in conjunction with their biodiversity information. Also in biodiversity field, DNA sequence data such as DNA barcoding has become essential as evidence for species identification and phylogenetic analysis as well as occurrence and morphological information. To accelerate biodiversity informatics, it is important to utilize both biodiversity occurrence and morphology data, and bioinformatics sequencing data. There are many databases for biodiversity domain such as GBIF (The Global Biodiversity Information Facility) for species occurrence records, EoL (The Encyclopedia of Life) as a knowledge base of all species, and BOLD (The Barcode of Life Data) for DNA barcoding data. In genomics science, molecular data involving DNA and protein sequences have been captured by the DNA Data Bank in Japan (DDBJ), the European Bioinformatics Institute (EBI, UK), and the National Center for Biotechnology Information (NCBI, US) under the International Nucleotide Sequence Database Collaboration (INSDC) for more than 30 years. Recently, NCBI launched a new database called BioCollections, including 7,930 culture collections, museums, herbaria, and other natural history collections. In addition, we can submit biodiversity information such as specimen voucher IDs, BOLD IDs, and latitude/longitude with DNA sequences. To find out the current situation, I downloaded GenBank (Nucleotide) files (updated at 22 Feb 2019) from the NCBI FTP (file transfer protocol) site and extracted biodiversity features including specimen voucher IDs and BOLD IDs. For Insecta, there are 2,427,343 sequence entries with specimen voucher ID and 1,766,142 entries with BOLD ID of 3,389,495 total entries. The most abundant species with voucher IDs is "*Cecidomyiidae* sp. BOLD−2016" (Diptera) (35,861 sequence entries). The most frequently referred voucher ID is "USNMENT00921257" (1510 sequence entries), indicating *Stenamma megamanni* (Hymenoptera, Formicidae, Myrmicinae). For flowering plants (Magnoliophyta), of 3,094,140 total entries, 1,109,420 sequence entries are assigned with voucher IDs and 73,409 entries with BOLD IDs. Additionally, 79,891 matK entries and 63,821 rbcL entries are submitted with voucher IDs, without BOLD IDs. I also retrieved BOLD data for Insecta and flowering plants. The 2,368,801 GenBank entries are referred from 4,176,481 BOLD total entries for Insecta, and the 259,245 GenBank entries from 345,706 BOLD entries for flowering plants. Some DNA barcoding data exist redundantly in BOLD database because BOLD imports sequences from NCBI submitted as DNA barcoding data in BOLD. These entries have different BOLD IDs but same BIN_URL is assigned. Recently, high-throughput sequencing technology, also called next-generation sequencing technology (NGS), has made a great impact in genomic science. Biodiversity researchers became to perform not only DNA barcoding but also RNA-Seq with NGS. NGS also accelerates museomics activity. NGS data are archived to the Sequence Read Archive (SRA) database, and sample information is described in BioSample database in INSDC. To utilize NGS data for biodiversity field, we will need to integrate such databases and other biodiversity databases. We, Database Center for Life Science, tackle to integrate life science data with Semantic Web technology. We held annual meetings to integrate life science data, called BioHackathons, in which researchers from all over the world participated. We began to RDFize BioSample data, but we should import existing schemes used in the biodiversity field including Darwin Core.

Keywords

museomics, DNA sequence,DNA barcoding, database, NGS, semantic web technology, RDF

Presenting author

Takeru Nakazato

Funding program

The Life Science Database Integration Project.

Conflicts of interest

The author has declared that no competing interest exists.