Knowledge Extraction from Specimen-Derived Data from GenBank to Enrich Biodiversity Information

Takeru Nakazato ‡

‡ Database Center for Life Science, Mishima, Japan

Corresponding author: Takeru Nakazato (nakazato.tkr@gmail.com)

Abstract

DNA barcoding and environmental DNA (eDNA) are increasing the need for the utilization of gene sequences in the field of biodiversity. GBIF (Global Biodiversity Information Facility) and GGBN (Global Genome Biodiversity Network) are taking action on the treatment of gene sequences in the field of biodiversity (Finstad et al. 2020). Gene sequences have been collected and published by INSDC (International Nucleotide Sequence Database Collaboration) for over 30 years (Arita et al. 2020). Biodiversity information has been collected using standards such as Darwin Core (Wieczorek et al. 2012), but INSDC gene sequences are stored in their own format. In the field of bioinformatics, researchers are also organizing the BioHackathon series, notably the NBDC/DBCLS BioHackathon and the spin-off Biohackathon Europe, to standardize data through the Semantic Web (Garcia Castro et al. 2021, Vos et al. 2020), but the linkage with biodiversity information has just begun.

In this study, as an example of linking gene sequence information with biodiversity information, I attempted to construct an infrastructure for knowledge extraction by utilising gene sequence entries derived from museum specimens from GenBank (Sayers et al. 2020). I have previously surveyed the BOLD (The Barcode of Life Data System) (Ratnasingham and Hebert 2007) IDs listed in GenBank (Nakazato 2020). I downloaded the fish and insect data from the GenBank FTP (file transfer protocol) site. Then I extracted the descriptions in the "specimen_voucher" field and obtained 749,627 (28% of the fish entries in GenBank) and 1,621,890 (13%) specimen IDs, respectively. I also extracted from the "note" field approximately 1000 entries describing the type of the specimen, such as "holotype", "lectotype", and "paratype". These extracts include descriptions written in natural language. NCBI (National Center for Biotechnology Information) publishes the BioCollections database (Sharma et al. 2019), and these data may be able to refine the description.

In the future, I plan to map these extracted IDs to the collection IDs in the biodiversity information database. This will enable us to enrich the biodiversity information with GenBank descriptions, for example, by adding articles listed in GenBank as references to the specimen data.

Keywords

RDF, linked open data, Wikidata, voucher specimen, natural language processing, taxonomic name

Presenting author

Takeru Nakazato

Presented at

TDWG 2021

Conflicts of interest

References

- Arita M, Karsch-Mizrachi I, Cochrane G (2020) The international nucleotide sequence database collaboration. Nucleic Acids Research 49 https://doi.org/10.1093/nar/gkaa967
- Finstad AG, Andersson A, Bissett A, Fossøy F, Grosjean M, Hope M, Kõljalg U, Lundin D, Nilsson H, Prager M, Jeppesen TS, Svenningsen C, Schigel D (2020) Publishing sequence-derived data through biodiversity data platforms. GBIF Secretariat https://doi.org/10.35035/doc-vf1a-nr22
- Garcia Castro LJ, Martin C, Lazarov G, Cernoskova D, Takatsuki T, Harrow J, Rebholz-Schuhmann D (2021) Measuring outcomes and impact from the BioHackathon Europe. BioHackrXiv https://doi.org/10.37044/osf.io/3dxhg
- Nakazato T (2020) Survey of Species Covered by DNA Barcoding Data in BOLD and GenBank for Integration of Data for Museomics. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59065
- Ratnasingham S, Hebert PN (2007) BARCODING: bold: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7 (3): 355-364. https://doi.org/10.1111/j.1471-8286.2007.01678.x
- Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, Karsch-Mizrachi I (2020) GenBank. Nucleic Acids Research 49 https://doi.org/10.1093/nar/gkaa1023
- Sharma S, Ciufo S, Starchenko E, Darji D, Chlumsky L, Karsch-Mizrachi I, Schoch CL (2019) The NCBI BioCollections Database. Database 2019 https://doi.org/10.1093/database/baz057

- Vos R, Katayama T, Mishima H, Kawano S, Kawashima S, Kim J, Moriya Y, Tokimatsu T, Yamaguchi A, Yamamoto Y, Wu H, Amstutz P, Antezana E, Aoki N, Arakawa K, Bolleman J, Bolton E, Bonnal RP, Bono H, Burger K, Chiba H, Cohen K, Deutsch E, Fernández-Breis J, Fu G, Fujisawa T, Fukushima A, García A, Goto N, Groza T, Hercus C, Hoehndorf R, Itaya K, Juty N, Kawashima T, Kim J, Kinjo A, Kotera M, Kozaki K, Kumagai S, Kushida T, Lütteke T, Matsubara M, Miyamoto J, Mohsen A, Mori H, Naito Y, Nakazato T, Nguyen-Xuan J, Nishida K, Nishida N, Nishide H, Ogishima S, Ohta T, Okuda S, Paten B, Perret J, Prathipati P, Prins P, Queralt-Rosinach N, Shinmachi D, Suzuki S, Tabata T, Takatsuki T, Taylor K, Thompson M, Uchiyama I, Vieira B, Wei C, Wilkinson M, Yamada I, Yamanaka R, Yoshitake K, Yoshizawa A, Dumontier M, Kosaki K, Takagi T (2020) BioHackathon 2015: Semantics of data for life sciences and reproducible research. F1000Research 9 https://doi.org/10.12688/f1000research.18236.1
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715