BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e26102
doi: 10.3897/biss.2.26102

OPEN ACCESS

Conference Abstract

A Challenge to Integrate Bioinformatics and Biodiversity Informatics Data as Museomics

Takeru Nakazato [‡]

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

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

Received: 23 Apr 2018 | Published: 22 May 2018

Citation: Nakazato T (2018) A Challenge to Integrate Bioinformatics and Biodiversity Informatics Data as Museomics. Biodiversity Information Science and Standards 2: e26102. https://doi.org/10.3897/biss.2.26102

Abstract

Museum-preserved samples are attracting attention as a rich resource for DNA studies. Museomics aims to link DNA sequence data back to the museum collection. Molecular biologists are interested in morphological information including body size, pattern, and colors, and sequence data have also become essential for biodiversity research as evidence for species identification and phylogenetic analysis.

For more than 30 years, molecular data, such as DNA and protein sequences, have been captured by the DNA Data Bank of 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). INSDC provides collected molecular data to researchers as public databases including GenBank for DNA sequences and Gene Expression Omnibus (GEO) for gene expression. These three institutes synchronize archived data and publish all data on an FTP (File Transfer Protocol) site so that it is available for big data analysis.

In recent years, high-throughput sequencing technology, also called next-generation sequencing (NGS) technology, has been widely utilized for molecular biology including genomics, transcriptomics, and metagenomics. Biodiversity researchers also focus on NGS data for DNA barcoding and phylogenetic analysis as well as molecular biology. Additionally, a portable NGS platform, MinION (Oxford Nanopore Technologies), has been launched, enabling biodiversity researchers to perform DNA sequencing in the field. Along with GenBank and GEO data, INSDC accepts NGS data and provides a public primary database, called the Sequence Read Archive (SRA). As of March 2018, 6.4 Peta Bases of NGS data is freely available under more than 130,000 projects in SRA. The Database Center for Life Science (DBCLS) provides a search engine for public NGS data, called DBCLS SRA (http://sra.dbcls.jp/) in collaboration with DDBJ. SRA contains not only raw sequence reads or processed data mapped to genome, but also information on the experimental design, including project types, sequencing platforms, and sample species. Researchers can use this data to refine their search results. We also linked publications referring to NGS data to the corresponding SRA entries.

The mission of DBCLS is to accelerate the accessibility of life science data. Collected data used to be described in the Excel-readable tabular format, but these formats are difficult to merge with other databases because of the ambiguity of labels. To overcome this difficulty, we recently integrated 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. UniProt and Ensembl databases currently provide an RDF (Resource Description Framework) version of curated genome and protein data, respectively. In the biodiversity domain, there are many databases 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 BoL (The Barcode of Life) for DNA barcoding data. RDF is utilized to describe Darwin Core based data so that bioinformatics and biodiversity informatics researchers can technically merge both types of data. Currently, specimen data and DNA sequence data are not linked. Museomics starts with cross-referencing specimen and sequence IDs and by making data sources comply with an existing standard.

Keywords

NGS, museomics, database, DNA sequence, museum collection

Presenting author

Takeru Nakazato

Funding program

The Life Science Database Integration Project.

Conflicts of interest

The author has declared that no competing interest exists.