Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D1M06T

Title: Protein Knowledge Integration Center (ProKiC)

Creator: Elisha Wood-charlson - **ORCID:** 0000-0001-9557-7715

Affiliation: Lawrence Berkeley National Laboratory (lbl.gov)

Principal Investigator: Valérie de Crécy-Lagard

Data Manager: Elisha Wood-Charlson

Funder: National Science Foundation (nsf.gov)

Funding opportunity number: NSF 22-608

Template: NSF-BIO: Biological Sciences

Project abstract:

Advances in DNA sequencing technology have provided biologists with the DNA sequences of over 200,000 organisms. Yet knowing the genetic makeup of these species is only the first step in understanding their biology in order to apply this knowledge to solve essential societal problems. The next steps are to understand the molecular and cellular biology of these organisms—how each of the proteins encoded in the genome functions to create cellular systems underlying life and enabling adaptation and interactions with other individuals and species. This critical step can only be made if a high fraction of proteins in any given sequenced organism is accurately linked to a function that can be integrated into a cellular network. Many individual and global initiatives have worked diligently in capturing this knowledge over the last two decades, but major bottlenecks remain. Indeed, on average only 50% of sequenced proteins can be accurately linked to a function for most species all over the tree of life. Some progress has been made, but protein function information is complex, often inaccessible, and unintegrated, preventing its widespread use and hampering scientific progress. Solving this problem requires a coordinated approach among disciplines that have traditionally been siloed The Protein Knowledge Integration Center (ProKiC) will bring together experimental biologists, biocurators (biological data scientists), and computational biologists to overcome barriers to the generation and dissemination of protein function information that have been impeding

progress in molecular and cellular biology.

Start date: 02-01-2024

End date: 01-31-2029

Last modified: 06-22-2023

Copyright information:

Protein Knowledge Integration Center (ProKiC)

Data and Materials Produced

Describe the types of data, physical samples or collections, software, curriculum materials, and other materials to be produced in the course of the project. (For collaborative proposals, the DMP must cover all the various data types being collected by each collaborator.)

ProKiC will not generate new data but will be curating and improving upon existing community data and knowledge. All updates to datasets will be versioned, fully provenanced and documented, and re-released to the source repository (e.g., UniProt, Gene Ontology Central). All software will be open source and available on GitHub. All teaching materials will be registered at the Network for Integration of Bioinformatics in Life Science Education (NIBLSE), a learning resource collection hosted on Quantitative Undergraduate Biology Education and Synthesis (QUBES), and interactive data science and curation modules will be available for immediate integration into coursework as part of the KBase Educators program.

Standards, Formats and Metadata

Describe the standards to be used for all the data types anticipated, including data or file format and metadata. [Note: Where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies.]

ProKiC will follow all community standards, as available and described in the proposal, such as Gene Ontology (GO) and others developed by the Open Biomedical Ontologies (OBO) Foundry. For genomic data and metadata, ProKiC will follow conventional data formats such as GFF (Genome Feature Format), GAF (Gene Association Format), and FASTA as appropriate. In addition, a significant component of the ProKiC effort will be to work with the community to develop and establish adoption of standards for generating and reporting protein functional annotations. Center activities will support training for the community around best practices in applying functional annotation standards and working with funders and publishers to disseminate and support adoption. Successful adoption will be demonstrated by functional annotation standards being reported in proposal DMPs, and in curated data available in public repositories (e.g., UniProt), and data analysis platforms (e.g., KBase) prior to the release of associated publications.

Roles and Responsibilities

Describe the roles and responsibilities of all parties with respect to the management of the data (including contingency plans for the departure of key personnel from the project).

The PI and the Open Science coordinator are responsible for clear and comprehensive guidance across the project with respect to data management, citation and credit for intellectual contributions, and transparent reporting of data releases. The cyberinfrastructure hub will be responsible for creating automated methods to support rapid release and accurate attribution of enhanced data products to the main repositories and open science platforms. All postdoctoral fellows, graduate students, and working group leads will be responsible for reviewing and following ProKiC guidelines and best practices for data management and attribution. At the end of their tenure with ProKiC, they will follow reporting guidelines to ensure knowledge transfer around data and materials generated from their projects.

Dissemination Methods

Describe the dissemination methods that will be used to make data and metadata available to others during the period of the award, and any modifications or additional technical information regarding data access after the grant ends.

All data releases will be fully described, documented, and published with a DOI via the KBase platform, which enable credit and attribution to the original providers of the data, as well as any the curation efforts contributed by the ProKiC team and community members, and citation of any software tools used to enhance/update the data. Releases will be contributed back (and versioned) into the repository of origin (described above), and communicated through the ProKiC news channels (website, newsletter, social media) and activities (monthly seminar series, conference events, etc.), as well as any news feeds or announcement channels available at the repository. Software will be available on GitHub, reproducible notebooks via Jupyter lab and KBase, and teaching materials at QUBES.

Policies for Data Sharing and Public Access

Describe the PI's policies for data sharing, public access and re-use, including re-distribution by others and the production of derivatives. Where appropriate, include provisions for protection of privacy, confidentiality, security, intellectual property rights and other rights.

All data products, training materials, and standards developed by ProKiC will be licensed for reuse without restriction (e.g., Creative Commons Zero, "no rights reserved"). The software will be licensed under an Open Source license, allowing reuse, modification, and redistribution (e.g., BSD 3.0 or Apache 2.0). No private or confidential materials will be generated, nor will any products be subject to intellectual property rights.

Archiving, Storage and Preservation

Where relevant, describe plans for archiving data, samples, software, and other research products, and for on-going access to these products through their lifecycle of usefulness to research and education. Consider which data (or research products) will be deposited for long-term access and where. (What physical and/or cyber resources and facilities (including third party resources) will be used to store and preserve the data after the grant ends?)

All data products and training materials will be publicly available via community repositories and data platforms and knowledge bases (e.g., UniProt and KBase) or, when this does not within the scope of these knowledge bases, into generic repositories such as Zenodo. KBase is led by Berkeley Lab and is managed by the University of California Regents. Through that partnership, all KBase data releases with DOIs will also be deposited at the generalist repository affiliated with the California Digital Library (currently Merritt). In addition, KBase does routine backups to the cloud for glacial storage, in the unlikely event of data loss in the system. UniProt is managed by the three members of the UniProt Consortium - the SIB Swiss Institute of Bioinformatics (Geneva, Switzerland), the European Bioinformatics Institute (Hinxton, UK), and the Protein Information Resource (PIR) (Georgetown, DC and Delaware). Software will be available via GitHub archives, and immutable version freezes will be available with a DOI in Zenodo. Version freezes of the software will ensure reproducibility of analyses since using versioned code will allow for consistency in software use.

Planned Research Outputs

Dataset - "ProKiC Annotation Update Release"

Software - "ProKiC Annotation Tools"

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
ProKiC Annotation Update Release	Dataset	Unspecified	Restricted	KBase UniProtKB		Creative Commons Zero v1.0 Universal	None specified	No	No
ProKiC Annotation Tools	Software	Unspecified	Open	github		Apache License 2.0	None specified	No	No