Plan Overview

A Data Management Plan created using DMPonline

Title: BioAIUnlock: Identification of stress-resistant anaerobic microorganisms to unleash the potential of lignocellulosic biomass in future biorefineries

Creator: Tong Liu

Principal Investigator: Tong Liu

Affiliation: Swedish University of Agricultural Sciences

Funder: FORMAS

Template: FORMAS Template

ORCID iD: 0000-0002-6456-4767

Project abstract:

In the rapidly changing international context, advancing Sweden towards a sustainable, fossil-free future is imperative. Anaerobic digestion (AD) of lignocellulosic biomass, key for biorefineries, enables the production of chemicals like fatty acids, alcohols, hydrogen, and biomethane. The potential of lignocellulosic materials for AD is underexploited due to their recalcitrant structure and low nutrient levels. Strategies like pre-treatment and co- digestion have been explored to boost AD efficiency but often introduce inhibitors like humic acids and ammonia, impacting microbial performance.Emerging research suggests microbial communities' capacity to adapt to these inhibitors, highlighting the importance of understanding microbial resilience for optimizing lignocellulose conversion. Enhancing work by the collaborated groups, this project introduces a novel Isotopic Meta-Omics and ML pipeline fusion to analyze AD microbial and enzyme communities, uncovering unique insights into stress tolerance mechanisms.This project aims to use these tools to identify resilient lignocellulolytic communities and enzymes, enhancing lignocellulose degradation through enrichment and bioaugmentation, in collaboration with industry. By deepening our understanding of AD processes and microbial adaptation, this project has great potential to improve biorefinery operations and strengthen collaborations, propelling Sweden closer to energy and chemicals independence and resilience.

ID: 202787

Start date: 01-01-2025

End date: 31-12-2028

Last modified: 22-04-2026

Grant number / URL: 2024-00510

Copyright information:

BioAIUnlock: Identification of stress-resistant anaerobic microorganisms to unleash the potential of lignocellulosic biomass in future biorefineries

GENERAL INFORMATION

Project Title

BioAIUnlock: Identification of stress-resistant anaerobic microorganisms to unleash the potential of lignocellulosic biomass in future biorefineries

Project Leader

Tong Liu; tong.liu@slu.se; Department of Molecular Sciences, SLU; https://orcid.org/0000-0002-6456-4767

Registration number/corresponding

2024-00510

Version

1.0

Date

20260422

DESCRIPTION OF DATA – REUSE OF EXISTING DATA AND/OR PRODUCTION OF NEW DATA

How will data be collected, created or reused?

a) Data will be generated from anaerobic digestion experiments (enrichment cultures, bioreactors) and meta-omics analyses (metagenomics, metatranscriptomics, protein-SIP). Sequencing will be performed using Illumina platforms. Bioinformatics processing will be conducted using established pipelines (e.g. SqueezeMeta), and downstream analyses and machine learning will be performed in R

and Python.

b) Public datasets from biogas systems may be reused for comparative analyses and model training. No major restrictions are expected beyond standard repository terms (e.g. citation and reuse conditions).

c) Data provenance will be documented through detailed metadata, including sample origin, experimental conditions, and processing steps. All analysis workflows, parameters, and software versions will be recorded and version-controlled.

d) Existing datasets will be reused where appropriate (e.g. for model development), but new experimental data are required due to the project’s focus on controlled conditions and specific stress-response mechanisms not available in public datasets

What types of data will be created and/or collected, in terms of data format and amount/volume of data?

e) The project will generate mixed data types, including:

* Numeric data (e.g. methane production, reactor parameters; spreadsheets/databases)
* Sequencing data (metagenomics, metatranscriptomics)
* Protein data (metaproteomics/protein-SIP)
* Text-based metadata and documentation

f) Data will be stored in standard formats, including:

* Sequencing data: FASTQ, FASTA
* Annotation tables: CSV/TSV
* Metadata and experimental data: CSV/XLSX
* Scripts and documentation: TXT, R, Python

g) Open and widely used formats are prioritised to ensure interoperability and long-term reuse. These formats are standard in bioinformatics workflows and compatible with public repositories (e.g. ENA/NCBI) and commonly used software at SLU.

h) Data volumes are expected to be moderate to large:

* Sequencing data: several terabytes (TB) across the project
* Processed datasets and metadata: gigabyte (GB) scale
* Scripts and documentation: small (MB scale)

DOCUMENTATION AND DATA QUALITY

How will the material be documented and described, with associated metadata relating to structure, standards and format for descriptions of the content, collection method, etc.?

a) Metadata will include sample identifiers, origin, experimental conditions (e.g. inhibitors, temperature), sequencing details, and analytical workflows, enabling data discovery and interpretation.

b) Community standards will be followed where applicable, including MIxS for sequencing data and commonly used bioinformatics annotation frameworks (e.g. KEGG, COG, PFAM).

c) Data will be organised in structured folders by work package and data type (raw, processed, results). Consistent naming conventions will be applied, and version control (e.g. Git) will be used for scripts and analysis workflows.

d) Additional documentation will include methodological descriptions, data processing steps, variable

definitions, and units of measurement to support reproducibility.

e) This information will be recorded in metadata tables (CSV/TSV), README files, and version-controlled scripts, with links between datasets and documentation to ensure traceability

How will data quality be safeguarded and documented (for example repeated measurements, validation of data input, etc.)?

Data quality will be ensured through standardized experimental protocols, use of biological replicates, and controlled laboratory conditions. Instrument calibration and routine checks will be performed where applicable.

Sequencing data will undergo quality filtering, trimming, and validation using established pipelines. Data consistency will be maintained through standardized metadata templates and controlled vocabularies.

All processing steps, parameters, and software versions will be documented, and analysis workflows will be version-controlled to ensure traceability and reproducibility

STORAGE AND BACKUP

How is storage and backup of data and metadata safeguarded during the research process?

Data and metadata will be stored on secure institutional servers at SLU with regular automated backups in accordance with institutional policies. Access will be restricted where appropriate to ensure data integrity.

These procedures follow SLU guidelines and established infrastructure for research data management. As SLU is a Swedish governmental higher education institution, data handling, archiving, and screening are conducted in compliance with national regulations and institutional requirements

How is data security and controlled access to data safeguarded, in relation to the handling of sensitive data and personal data, for example?

The project does not involve sensitive or personal data. Data security will be ensured through storage on secure SLU servers with controlled access and regular backups.

Data handling follows SLU policies and established institutional procedures. As SLU is a Swedish governmental higher education institution, data management, access control, and archiving comply with national regulations and institutional requirements.

LEGAL AND ETHICAL ASPECTS

How is data handling according to legal requirements safeguarded, e.g. in terms of handling of personal data, confidentiality and intellectual property rights?

The project does not involve personal data or sensitive information; therefore, requirements related to GDPR, informed consent, anonymisation, or restricted access are not applicable.

All data will be handled in accordance with SLU policies and Swedish legislation. Data will be securely stored with controlled access during the project and made openly available upon publication where possible.

Intellectual property rights will follow SLU regulations. Data sharing and reuse will be supported through open repositories and appropriate licensing, while any collaboration-related data use will be governed by formal agreements if needed

How is correct data handling according to ethical aspects safeguarded?

No ethical issues affecting data handling are expected, as the project does not involve human subjects, personal data, or sensitive information. Data will be stored, transferred, and shared using secure institutional systems at SLU, with controlled access during the project and open access upon publication where appropriate.

The project will follow good scientific practice and comply with SLU guidelines and relevant national and international standards for research integrity. Ethical review is not required for this type of research.

ACCESSIBILITY AND LONG-TERM STORAGE

How, when and where will research data or information about data (metadata) be made accessible? Are there any conditions, embargoes and limitations on the access to and reuse of data to be considered?

a) Data will be made discoverable and shared through established repositories. Sequencing data will be deposited in ENA/NCBI, while processed data and code will be shared via Zenodo or GitHub. Metadata will ensure indexing and discoverability.

b) Data storage and archiving will follow SLU procedures and national requirements for Swedish higher education institutions.

c) Data and metadata will be made publicly available upon publication or at the latest by the end of the project. Short embargo periods may apply to allow publication or protection of potential intellectual property.

d) Data will be openly accessible to the research community. If necessary, temporary restrictions may be applied, but efforts will be made to minimise limitations and ensure broad reuse

In what way is long-term storage safeguarded, and by whom? How will the selection of data for long-term storage be made?

Long-term storage will be ensured through deposition in established public repositories (e.g. ENA/NCBI, Zenodo) and institutional storage systems at SLU, which provide secure archiving and long-term access. The project leader (Tong Liu, SLU) is responsible for data selection and preservation.

Data selected for long-term storage will include raw sequencing data, processed datasets, metadata,

and scripts necessary to reproduce results, while redundant or intermediate files will not be retained.

These procedures follow SLU guidelines and national requirements for Swedish governmental higher education institutions regarding archiving and screening. The data are expected to be reused by the scientific community for microbial ecology, bioinformatics, and biorefinery-related research

Will specific systems, software, source code or other types of services be necessary in order to understand, partake of or use/analyse data in the long term?

Data can be accessed and reused using standard, widely adopted tools. Sequencing data (FASTQ/FASTA) can be handled with common bioinformatics software, and downstream analyses can be performed in R and Python. No proprietary or specialised software is required, ensuring long-term accessibility.

All scripts and workflows will be shared and documented to support reproducibility. Data will be deposited in established repositories (e.g. ENA/NCBI, Zenodo), where access and data requests are managed through the repository systems

How will the use of unique and persistent identifiers, such as a Digital Object Identifier (DOI), be safeguarded?

Persistent identifiers will be ensured by depositing data in established repositories (e.g. ENA/NCBI, Zenodo), which automatically assign accession numbers and DOIs. These identifiers enable reliable citation, tracking, and long-term accessibility.

The data are expected to be reused in microbial ecology, bioinformatics, and biorefinery research. Use of persistent identifiers will facilitate discovery, integration with other datasets, and reproducibility across studies

RESPONSIBILITY AND RESOURCES

Who is responsible for data management and (possibly) supports the work with this while the research project is in progress? Who is responsible for data management, ongoing management and long-term storage after the research project has ended?

The project leader, Dr. Tong Liu (SLU), is responsible for overall data management, including data collection, metadata generation, quality control, storage, backup, and data sharing. Collaborators are responsible for managing the data they generate, following agreed procedures under the coordination of the project leader.

During the project, data will be managed using SLU infrastructure. After project completion, long-term storage and archiving will be ensured through public repositories (e.g. ENA/NCBI, Zenodo) and institutional systems at SLU.

The project leader is responsible for implementing, monitoring, and updating the DMP throughout the project. The DMP will be reviewed regularly and revised if needed to reflect changes in data handling practices

What resources (costs, labour input or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)? What resources will be needed to ensure that data fulfil the FAIR principles?

Data management will be integrated into the project and primarily covered by the project leader, Dr. Tong Liu (SLU), who is responsible for data capture, metadata production, quality control, storage, backup, archiving, and sharing. Collaborators will manage the data they generate under coordinated procedures.

Resources include institutional storage systems at SLU, national e-infrastructure (e.g. NAISS/SciLifeLab), and high-performance computing for bioinformatics and machine learning analyses. Standard tools (R/Python, Git) will support data processing, documentation, and version control.

Ensuring FAIR data will rely on the use of standardized metadata, open formats, and deposition in public repositories (e.g. ENA/NCBI, Zenodo), with persistent identifiers (DOI/accession numbers).

All costs related to storage, computing, and data management are covered within the project’s existing budget and infrastructure, and no separate dedicated resources are required