Combined computing and experimental platform for microbial strain optimization towards biochemicals production

A Data Management Plan created using DMPonline.be

Creator: Meesha Katyal

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

ID: 206164

Start date: 01-11-2023

End date: 31-10-2027

Project abstract:

Microbes are increasingly being used as cell factories to produce chemicals in a sustainable way that does not rely on fossil fuels. However, the chemical of interest or intermediates in its production pathway often pose a significant metabolic burden on the host cell and in many cases exert toxic effects, thereby preventing microbial end-product from reaching high concentrations. To circumvent these issues, we propose an integrated approach that combines metabolic and tolerance engineering to optimize microbial cell factories. More specifically, we set out to gain insight into the factors that impede microbial production of chemicals using 1,2-propanediol (1,2-PDO) production in *Escherichia coli* as a proof of concept. First, the impact of 1,2-PDO overproduction on the entire *E. coli* metabolome will be investigated using a combination of computational tools, 13C-tracing experiments, and energy- and cofactor-sensitive reporters. Second, the microbial (adaptive) stress response towards the 1,2-PDO's precursor lactaldehyde will be elucidated using CRISPRi screening and adaptive laboratory evolution. Finally, gathered insights at the metabolic and tolerance level will be combined to create an optimized 1,2-PDO overproducing *E. coli*. The latter will demonstrate that a combined metabolome-remodelling and tolerance-improving framework holds great potential to maximize microbial productivity alcoholic and diol chemicals.

Last modified: 22-04-2024

Combined computing and experimental platform for microbial strain optimization towards biochemicals production

FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	**Description**	**New or reused**	**Digital or Physical**	**Digital Data Type**	**Digital Data format**	**Digital data volume (MB/GB/TB)**	**Physical volume**
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Primers	Primer pairs designed to create genomic mutants. Used in all work packages.	Generate new data	Digital	Experimental	.csv	<100MB	
Primers	Primer pairs designed to create genomic mutants. Used in all work packages.	Generate new data	Physical	Experimental			Vials stored at -20°C
Genomic mutant strains	Deletion mutant *E. coli* strains created using CRISPR-FRT	Generate new data	Physical	Experimental			Glycerol stocks frozen at -80°C in 2mL cryotubes
DNA sequences	Sanger sequencing to confirm that primers and the mutant strain sequences	Generate new data	Digital	Experimental	.fasta	<1GB	
DNA gel images	Images of DNA gel electrophoresis	Generate new data	Digital	Experimental	.tiff	<1GB	

DNA oligonucleotides	DNA oligos generated for the construction of mutant strains	Generate new data	Physical	Experimental			Vials stored at -20°C
CRISPRi-Seq pools	Plasmid stocks and transformed gRNA libraries	Generate new data	Physical	Experimental			Glycerol stocks frozen at -80°C in 2mL cryotubes
NGS sequencing data	Next-generation sequencing CRISPRi libraries created with a 2-step PCR	Generate new data	Physical	Experimental			Vials stored at -20°C
QC data	Quality control data generated from nanodrop, qubit and size analyses	Generate new data	Digital	Experimental	.txt, .csv, .xlsx	<1GB	
NGS sequencing data	Millions of short read sequences, per sample, generated by NGS machinary at core facilities (or in-house).	Generate new data	Digital	Experimental	.fastq (or its zipped versions)	<5TB	
Reference sequence data	Reference sequences that can be obtained from NIH / Pubmeb or other publicly available databases.	Reuse existing data	Digital	Observational	.fasta	<100MB	
Differentially targetted genes and GO enrichments	Analysis of the NGS data will result in a list of fold changes and p-values for each sample.	Generate new data	Digital	Experimental	.txt, .csv, .xlsx	<100MB	
Cytoflex data	Cell fluorescent measurements using cytoflex	Generate new data	Digital	Observational	.xit, .fcs, .xml	<100GB	
MIC data	Lactaldehyde minimal inhibitory concentration for the wild-type and mutant strains	Generate new data	Digital	Experimental	.csv, .xlsx	<100GB	

Growth data	Growth dynamics base on automatic OD measurements of wildtype and mutant strains on different carbon sources in a 96-well plate format	Generate new data	Digital	Experimental	.csv, .xlsx	<100GB	
Production efficiencies	Detection of metabolites using Echo-MS, ESI-MS or HPLC	Generate new data	Digital	Experimental	.xlsx, .pdf	<100GB	
WGS	Whole genome sequencing of all the mutant strains, evolved strains	Generate new data	Digital	Experimental	.fastq	<100GB	
Evolution of mutant strains	Monitoring growth while registering the evolution of lactaldehyde tolerance over time	Generate new data	Digital	Experimental	.xlsx, .cvs, .nd	<100GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

Two fluorescent biosensors were requested from another lab based on the paper: Development of a Metabolite Sensor for High-Throughput Detection of Aldehydes in *Escherichia Coli* (DOI: 10.3389/fbioe.2018.00118). An MTA was proofread and approved by VIB Innovation & Business officers.

Mutant strains will also be created based on the paper: Metabolic engineering of *Escherichia coli* for the de novo stereospecific biosynthesis of 1,2-propanediol through lactic acid (DOI: 10.1016/j.mec.2018.e00082).

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

Potential tech transfer will be discussed with the research and development offices of KU Leuven and VIB.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

Materials requested from other labs (e.g. biosensors) are subject to MTAs. This will be done in consultation with our host institutions' legal departments to minimize restrictions on the use of these materials.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- Yes

Jan Michiels holds the copyright of all the generated data while VIB and KU Leuven have the ownership. This is in accordance with the framework agreement between both the institutes.

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

For experimental data: All files are named using "yearmonthday_titleordescription" formal. The data is generated using standardized protocols saved in a OneNote notebook.
To keep the data organized and understandable, every day activities including protocols followed as well as brief results are registered in a OneNote sectionv as digital labnotes. Each page in the OneNote section stands for each month denoted and the experiments are noted per day while being labelled properly. At the end of this project, a readme.txt file will also be created to guide future students through the information recorded and how to navigate it.

For biological data: All bacterial samples are labelled with an identifier (eg strain name_plasmid_antibiotic) and stored in 2mL cryotubes at -80 °C. All the information including their location, sample name, selection markers etc is also noted down in a shared excel sheet.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

Some data types already have a standard metadata containing all the technical information that tells one about the settings, for example .fasrq NGS files. These files will be preserved with their original metadata file during the project. For other formats such as .txt, .csv, .xlsx files, section explanations or even an additional tab will be created to explain the data. For any other type of data where a standard metadata is not available, a readme.txt file will be created to explain all the data files as well as explain the analysis according to Dublin Core Metadata.

3. Data storage & back-up during the research project

Where will the data be stored?

Biological samples: Cryotubes will be stored in -80° freezers with restricted access.
Experimental results: All data will be stored on at least two locations, one of which will include a secure university device on which the data was generated and the other on a OneDrive location which is shared with the promoter and co-promoters.

How will the data be backed up?

Biological samples: A backup of selected important strains will be stored in a different location with restricted access
Experimental results: The use of Onedrive is backed up via a local copy and another back-up copy in the cloud / on the KU Leuven network drive. These different copies offer version control as well as allow easy recovery in case of data loss.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

Biological samples: There is sufficient storage space in the -80°C freezers available at Michiels lab.

Experimental results : OneDrive at KU Leuven offers 5 TB of data per user. In case of a shortage, network storage is purchased whenever needed.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Biological samples: Unauthorized persons do not have the access to the collection of strains stored in the lab
Experimental results: OneDrive storage is linked to my KU Leuven account which is secured by two-factor authentication. Its password is also changed on a regular basis. Rest of the data is stored on the university's secure network drives.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

Biological samples: The cost of -80°C freezers that are present in the Michiels lab are covered by general lab expenses.
Experimental results: The extra cost for any large volume storage will then be covered by general lab funding.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

Biological samples: All the strains will be stored in the -80 °C freezers present in the Michiels lab for at least 10 years after the end of this project, complying with the 10 year data preservation policy of the KU Leuven.
Experimental results: All the data and results will be stored on KU Leuven's servers for at least 10 years after the end of this project, complying with the 10 year data preservation policy of the KU Leuven. This will also follow automatic back-up procedures and the costs will be covered by the overhead budgets of KU Leuven.

Where will these data be archived (stored and curated for the long-term)?

Biological samples: The -80°C freezers present in the Michiels lab will store the strains for at least 10 years.
Experimental results: KU Leuven's servers will store the experimental data for 10 years. Published results will be deposited according to the

respective journal's policies.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Biological samples: The -80°C freezers present in the Michiels lab are run on general lab expenses.
Experimental results: The estimated cost of storage on the KU Leuven servers is upto 100 euros per year.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

We plan to communicate our results in top journals that require full disclosure upon the publication of all data. However, depending on the journal, accessibility restrictions may apply but the proper links to these data sets will always be provided in the corresponding publications.

If access is restricted, please specify who will be able to access the data and under what conditions.

No.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

All the published data will be available as per the guidelines of the journal. We aim to publish in journals which have an open access policy. The data that has not been published will be under an embargo period of 5 years after which it will be made public using the Research Data Repository (RaDaR), which is managed by KU Leuven.

When will the data be made available?

For the published results, they will be available soon thereafter according to the journal policies. For the unpublished results, the data will be made available after an embargo periods of 5 years.

Which data usage licenses are you going to provide? If none, please explain why.

Open Access publications in peer-reviewed journals are generally covered by a Creative
Commons Attribution License (CC-BY). The RaDaR repository also allows data sharing using the CC-BY license.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

A unique DOI will be assigned to the manuscript upon the publication. Similarly, the datasets will also receive a PID when they are deposited on an online platform.

What are the expected costs for data sharing? How will these costs be covered?

Biological samples: The shipment is generally paid for by the parties that request the samples.
Experimental results: The cost of publication is paid by the general lab expenses and an online repository is free of charge. Network storage charges are also covered by the general lab expenses.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Meesha Katyal

Who will manage data storage and backup during the research project?

Meesha Katyal

Who will manage data preservation and sharing?

Jan Michiels

Who will update and implement this DMP?

Jan Michiels