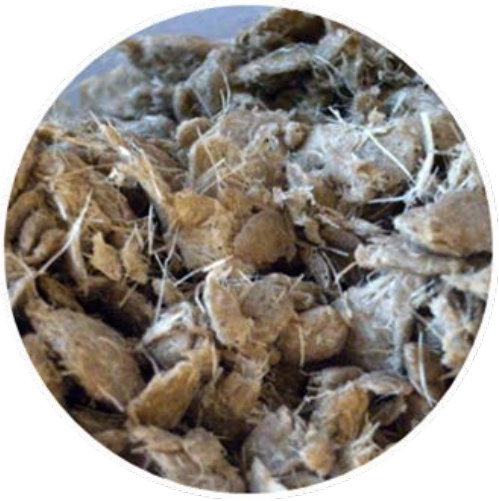
DD-DeCaF Project Grant Agreement n°686070

D1.1
Data Management Plan

31.08.2016

Dissemination level: Public

Open Research Data Pilot



About the DD-DeCaF project

The DD-DeCaF project - Bioinformatics Services for Data-Driven Design of Cell Factories and Communities - aims at enabling the use of a broad spectrum of omics data for the development of novel biotechnological applications by constructing a web-based bioinformatics platform for data-driven cell factory and microbial community design. Further information about the project and the partners involved are available under www.dd-decaf.eu.

Project coordinator

Project partners

About this document

This report corresponds to deliverable D1 of the DD-DeCaF project – Data Management Report. It has been prepared by:

Danish Technical University
Anker Engelunds Vej 1, 2800 Kongens Lyngby, DK
Markus Herrgard, Coordinator
E-mail: herrgard@biosustain.dtu.dk
Phone: +45 24 92 17 80

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 686070. The responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

Table of contents

1 Initial Data Management Plan

This is the first version of the Data Management Plan following the guidelines for the Open Research Data Pilot.

2 Data Summary

2.1 Purpose of the data collection

Data is collected by two end-user company partners (DSM and Biosyntia) in order to provide real-life data to validate the underlying data analysis/interpretation methods, developed in this project and to provide test cases for the full DeCaF platform including data visualization.

2.2 Relation to the objectives of the project

End-user validation and feedback is crucial for the purpose of developing a broadly usable cell factory and community design platform. End-user engagement is best achieved by applying the platform and underlying methods to data generated by the end-user companies in their own research and development projects.

2.3 Types and formats of data collected

The primary large-scale experimental data types collected are:

Genomics: Data collection is achieved by short read sequencing on Illumina sequencer platforms creating raw reads in fastq format. Reads are aligned to a reference genome in order to allow identification of genetic variants in production strains.

Transcriptomics: Data collection is achieved through RNA-seq on Illumina sequencer platforms creating raw reads in fastq format. These raw reads are processed into derived tabular data formats that consist of unique transcript identifies and the corresponding absolute RNA expression level in a particular condition.

Proteomics: Data collection is achieved through standard massspectrometry proteomics platforms (vendor depends on the partner) each of which produces its own proprietary file type. These files will be converted to the standard mzML format that allows deposition to public databases. The processed data is in tabular data format that consist of unique protein and peptide identifiers and the corresponding relative protein expression level in a particular condition.

Metabolomics: Data collection is achieved through mass spectrometry (LC/MS) and HPLC platforms. The raw data is not of primary interest as the file formats are

proprietary to each instrument vendor. Instrument vendor provided software will be used to obtain abslute metabolite concentration data in a tabular format that consist of unique metabolite identifiers and the corresponding absolute metabolite concentration in a particular condition.

Fluxomics: Data collection is achieved through mass spectrometry (GC/MS) and HPLC platforms. As above the raw data is not of interest, but from the raw data a series of derived data types will be generated including isotopomer distributions and final metabolic flux estimates. All the derived data can be represented in tabular formats consisting of unique identifiers (e.g. the metabolic reaction specified by the reactant and product metabolite ids and stoichiometry) and the corresponding measured/estimated data.

In addition to experimental data, this project will include creating, improving and extending genome-scale models of cellullar processes.These models will be stored in the SBML (Systems Biology Markup Language) format, which will allow importing the models to public domain model repositories (BioModels at EBI). We will follow MIRIAM (Minimum Information Required in the Annotation of Models) standards to ensure that identifiers used in the models are consistent with identifiers used with data (e.g. gene/protein/metabolite identifiers). During Model versioning during the project will be done by maintaining the models on github. Upon publication, the final versions of the models will be deposited in BioModels.

2.4 Existing data being re-used

Existing reference genomics, transcriptomics, proteomics, metabolomics and fluxomics data may be reused for wild type (i.e. non-engineered) strains in order to validate the quality of the data produced within the project. However, new data for wild-type strains will also be produced during the project in order to ensure that a consistent dataset is overall generated for all the strains studied.

Existing public domain metagenomics data will be used for mining of novel enzyme functions and for functionally annotating metagenomic datasets by partners EMBL and biobyte. These data are also available in public metagenome repositories, but with minimal annotation. There are no restrictions for the use of the public domain metagenomics data and the derived data generated from this data can also be freely shared. The DD-DeCaF project will not generate any additional metagenomics data.

Existing public domain genome-scale models will be used as basis for improvement or extension within the project. The majority of these models will be available free of restrictions, but a few of the models will come with restrictions for commercial use. These restrictions will be propagated to the derived models as required by the licenses attached to the original models.

2.5 Origin of the data

The data will originate from two end user partners who will provide the raw data in standardized formats (for transcriptomics and proteomics) and the processed data in tabular format (for all omics data types).

The models will originate from four of the academic partners (DTU, EMBL, Chalmers and EPFL). DTU partner will be responsible for ensuring that the models are in formats that allow reuse and integration of experimental data with the models.

2.6 Expected size of the data

The raw data for each genomics data set (one strain) is approximately 0.25 Gb.

The raw data for each transcriptomics data set (one strain, condition, replicate) is approximately 0.5 Gb.

The raw data for each proteomics data set (one strain, condition, replicate) is approximately 2 Gb.

Each of the processed tabular data sets (all omics data types) is of the order of 0.5 Mb.

The total number of individual data sets is approximately 20 strains/conditions x 3 replicates = 60 samples total for all other omics data types except genomics and 20 samples for genomics (no replicates needed). This gives a total experimental data volume of 160 Gb primarily consisting of proteomics and transcriptomics raw data.

The genome-scale models will each take few tens of Mb of space and don't represent a major data management challenge.

2.7 Data utility

The experimental data will be useful within the project for testing and validating methods and for testing features of the overall data analysis and visualization platform. Outside of the project, the data will be useful as reference data for cell factory design in yeast and E. coli and as part of compendia of omics datasets for these organisms.

The genome-scale models generated within the project will be useful for cell factory design as well as a number of other applications related to metabolic physiology both within and outside of the project.

3 FAIR data

3.1 Making data findable, including provisions for metadata

There are two types of metadata, that need to accompany the primary omics data collected within the project: 1) Metadata describing the general experimental setup (e.g. strain genotype, cultivation conditions, sampling time points) and 2) metadata describing the process of going from a particular microbial culture to the raw data (e.g. sampling, sample processing, relevant instrument settings). The metadata within this project will be collected in the ISA-Tab format, which allows metadata submission together with raw/processed data submission to relevant databases (e.g. those maintained by the European Bioinformatics Institute, EBI). We will also utilize the ISA-Tab format as a metadata exchange format with the data repositories built within the project. Unique dataset identifiers will be generated in conjunction with submission of the genomics, proteomics and transcriptomics data to the public domain databases (ENA, PRIDE and ArrayExpress at EBI respectively). There are not currently comparable metabolomics or fluxomics databases that would work for the type of data generated within this project. For metabolomics and fluxomics the data will be submitted to the general purpose research materials sharing platform Zenodo. These platforms create the necessary dataset IDs and DOIs for all the materials. In addition to public domain databases, we will also build a database within this project that will contain the specific omics data and metadata generated within this project in order to demonstrate the use of the data within the platform and allow easy use of the data by academic and industrial partners working on data analysis and visualization methods.

3.2 Making data openly accessible

All data generated within the project will be made openly accessible after an embargo period of maximum of two years from the generation of the dataset during which the data will only be accessible within the consortium to the partners that require data access for method or tool development. During the embargo period the data will be made available to the partners through the DeCaF platform developed in this project. After the embargo period, the data will be made available publicly both through the platform developed during the project and through public data repositories as described above. The embargo period will end upon public disclosure of the data in the form of a preprint or conference/journal article if this happens in less than 2 years from the generation of the data. The project will develop the software tools and APIs that allow accessing the data within the DeCaF platform. Public data repositories already provide such tools.

The genome-scale models developed during the project will be subject to the usual embargo period where they will be available only to consortium partners until publication (either preprint or conference/journal article). After publication, models will

be publicly available both through the DeCaF platform and through the BioModels database. Software tools developed by academic partners during the DD-DeCaF project will be available open source through the github repository. SME partners in the project may release some of their code open source, but also reserve the right to maintain proprietary code bases where it is deemed to be necessary for commercial reasons.

3.3 Making data interoperable

The DD-DeCaF project will create genomics, transcriptomics, proteomics, metabolomics and fluxomics data for two organisms - Escherichia coli and Saccharomyces cerevisiae. We will use standard gene, transcript and protein identifiers that are specified by the reference databases for these organisms (EcoCyc and SGD respectively). For metabolites we will use universal unique chemical identifier (InChI) that allow conversion to other types of commonly used identifiers such as SMILES, ChEBI, PubChem, CAS. Within the DeCaF platform we will also provide genome-scale models that utilize these same standard identifiers (MIRIAM) so that the data can be directly mapped to the models and used with the methods developed within the DD-DeCaF project. The models will be made available in the standard SBML format facilitating use of the models with different types of modeling and visualization software. We will use standard raw and processed data formats where possible (genomics, transcriptomics and proteomics) as outlined earlier in the document.

3.4 Increase data re-use

All omics data and associated metadata will after the embargo period (described above, maximum of two years from data generation or upon publication) be freely usable without restrictions. The embargo period will be to allow seeking patents or to prepare publications. No data will be generated for commercially sensitive strains or processes within this project. Data will be quality controlled during initial data analysis within the project, and poor quality data (with quality standards dependent on the type of data) will be discarded and new data will be generated.

The modified and extended genome-scale models will be made publicly available upon scientific publication at the latest. Models will be made available with the same licensing restrictions that apply to underlying models that are used as starting point. All models will be free to use and modify for non-commercial use, but some may require commercial use licenses. Within the DeCaF platform these licensing restrictions will be made explicit. Models will be quality controlled by verifying their predictive performance against standard publicly available benchmark datasets for each organism if these are available.

4 Allocation of resources

4.1 Estimation of the costs for making data FAIR

Since this project is focused on building a data analysis and processing platform with the goal of designing new cell factories and communities, the costs of making data and models FAIR are already included in the proposed work and no additional costs will be involved in this process. The project includes components where data deposition and sharing tools are developed.

4.2 Responsibilities for data management

Data management of data generated within the project is primarily handled by the coordinating partner DTU. DTU will develop the data management, analysis and visualization platform that is used within the project. Partners EMBL and biobyte will handle development of the metagenomic data mining platform, but this platform only uses existing public metagenomics data.

4.3 Costs and potential value of long term preservation

Long term preservation of the primary data and metadata is ensured by deposition of the data to public domain repositories (ENA, ArrayExpress, PRIDE, BioModels and Zenodo). The DeCaF platform developed within the project will also be maintained long term using internal resources available within the Novo Nordisk Foundation Center for Biosustainability at DTU.

5 Data security

During the embargo period the data and models will be stored in the data and model repositories developed within the DD-DeCaF project as part of the DeCaF platform. These repositories will include as a feature the possibility of restricting data and model sharing to specific partners or making them entirely public. Access control is handled by the REMS system developed by CSC. The DeCaF platform will include backup features for all the data as well as long term archiving (10 years). The same apply to the genome-scale models.

Deposition to public domain databases will be only done after embargo period is complete. These databases handle backups and archiving internally and are expected to provide very long term data storage.

6 Ethical aspects

The data generated in this project is for microbial cell factories and involves no human subjects.

7 Other

DTU has drafted a research data management policy that is aligned with the EC Horizon 2020 data management policy requirements. DD-DeCaF will follow the DTU policy once it is made official within the following months.