Project Number: **ANR-19-CE45-0021 / DFG 431572533**

Project Acronym: **MetClassNet**

New approaches to bridge the gap between genome-scale metabolic networks and untargeted metabolomics

DATA MANAGEMENT PLAN

Changelog

v0.1 – 01.05.2020	Initial release of DMP.

1. Data Summary

Metabolism is a key biological process which is modulated in living organisms in response to environmental exposure, genetic variations and diet. Metabolism involves thousands of small molecules (metabolites) connected by thousands of biochemical reactions. Together they form the metabolic network, which is represented *in silico* by a Genome Scale Metabolic Network (GSMN). The measurement of metabolites and their abundances is called metabolomics, which uses advanced analytical chemistry tools such as NMR or MS.

MetClassNet is proposing a new computational framework and novel methods to help tackle metabolomics challenges in data analysis and data interpretation by integrating information experimentally derived information and GSMNs using direct mapping, ontologies and chemical class information. Within MetClassNet, we showcase the benefit of the computational framework with data sets obtained in the context of studies related to ageing, toxicology, cancer and nutrition.

Experimental LC/MS metabolomics data from three different organisms and application areas are used in **WP4 Biological Application and Validation**. Here we prepare datasets for framework development in **Task 4.1**, and generate new (reference) data sets in **Task 4.2**. The Following data sets will be used:

Human datasets:

Data Descriptor MTBLS17 “Utilization of Metabolomics to Identify Serum Biomarkers for Hepatocellular Carcinoma in Patients with Liver Cirrhosis”

Author(s): Ressom HW, Xiao JF, Tuli L, Varghese RS, Zhou B, Tsai TH, Ranjbar MR, Zhao Y, Wang J, Di Poto C, Cheema AK
Organism(s): Homo-Sapiens (Serum)

Data Descriptor MTBLS28 “Non-invasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer”

Author(s): Ewy Mathe, Andrew D Patterson, Majda Haznadar, Soumen K Manna, Kristopher W Krausz, Elise D Bowman, Peter G Shields, Jeffrey R Idle, Philip B Smith, Anami Katsuhiro, Dickran G Kazandjian, Frank J Gonzalez, Curtis C Harris
Organism(s): Homo-Sapiens (Urine)

Data Descriptor MTBLS737 “Metabolic Footprinting of a Clear Cell Renal Cell Carcinoma in vitro Model for Human Kidney Cancer Detection”

Author(s): Knott ME, Manzi M, Zabalegui N, Salazar MO, Puricelli LI, Monge ME
Organism(s): Homo-Sapiens (Cell line + Serum)

Data Descriptor MTBLS401 “Mass spectrometry based metabolomics for in vitro systems pharmacology: Pitfalls, challenges, and computational solutions”

Author(s): Stephanie Herman, Mats Gustafsson, Kim Kultima, Payam Emami Khoonsari, Obaid Aftab, Shibu Krishnan, Emil Strömbom, Rolf Larsson, Ulf Hammerling, Ola Spjuth
Organism(s): Homo-Sapiens (Cell lysate)

Data Descriptor (Metabolomics Workbench) M86G6V “Lung Cancer Plasma Discovery”

Organism(s): Homo-Sapiens (Plasma/Serum)

Data Descriptor (Metabolomics Workbench) M8FG61 “Colorectal Cancer Detection Using Targeted Serum Metabolic Profiling”

Organism(s): Homo-Sapiens (Serum)

Data Descriptor (Metabolomics Workbench) M80018 “Metabolic Profiling of Visceral and Subcutaneous Adipose Tissue from Colorectal Cancer Patients: Unraveling the Link Between Adipose Tissue and Cancer”

Organism(s): Homo-Sapiens (Adipose Tissue / Serum)

Plant datasets:

Data Descriptor MTBLS 1582 “Phytophthora Infection of Tomato”

Author(s): Micha Gracianna Devi, Ulrike Smolka, Steffen Neumann, Sabine Rosahl and Gerd Balcke
Organism(s): Solanum habrochaites, Solanum lycopersicum

Caenorhabditis elegans datasets:

Data Descriptor MTBLS1587 “HLH-30 dependent rewiring of metabolism during starvation in C. elegans (metabolomics)”

Author(s): Kathrine B. Dall, Jesper F. Havelund, Eva Bang Harvald, Michael Witting, Nils J. Færgeman
Organism(s): Caenorhabditis elegans

Data Descriptor MTBLS1586 “HLH-30 dependent rewiring of metabolism during starvation in C. elegans (lipidomics)”

Author(s): Kathrine B. Dall, Jesper F. Havelund, Eva Bang Harvald, Michael Witting, Nils J. Færgeman

Organism(s): Caenorhabditis elegans

For use in the MetClassNet framework, we will require and prepare all data to be compliant with the MetaboLights repository guidelines. Additionally, we will use the PSI data standard mzTab-M and the non-proprietary spectral library formats MSP or MGF for processed data.

Data analysis will be performed in the context of **Genome Scale Metabolic Networks (GSMN)**. We will use previously published or prepared GSMN models in the formats supported by the MetExplore system, in particular Systems Biology Markup Language (SBML).

Experimental networks, like correlation networks, mass-difference networks will be formatted using text- or XML based graph formats such as GML.

Chemical knowledge will be used from chemical ontologies such as ChemOnt and ChEBI in the Open Biology Ontology (OBO) or the Ontology Web Language (OWL) formats. Links between the GSMN and chemical ontologies will be via the InChIkey hash value or chemical class identifiers.

Data acquired within the projects and existing data sets improved to the requirements of our computational framework can be used by bioinformaticians involved in computational metabolomics developments to compare their methods. The data sets can also serve as educational material, how the reusability can be maximised. Finally, the acquired data also increases the publicly available studies related to ageing, toxicology, cancer and nutrition.

All partners agreed to pursue an open science approach, where programs, data and samples are transparently available to all partners, and eventually to the public as soon as possible. Members of the respective laboratories will have full access to the resources (datasets, bioinformatic analyses, etc.). All software developed in the project will be published under an accepted Open Source license, and available through one of the common source code repositories (e.g. Github, Bitbucket etc.). All data acquired and used will be available at public research data repositories, such as MetaboLights. For reproducible (re)analysis all scripts will be organized in Jupyter notebooks and/or workflow engines like KNIME or Galaxy.

2. FAIR data

2.1. Making data findable, including provisions for metadata

All datasets created under the MetClassNet umbrella will be submitted to MetaboLights initially as private studies, and metadata improved until it complies with MetaboLights requirements. Sensitive data, particularly the ones from Human studies, can not be shared publicly via MetaboLights without full explicit consent, see Ethics section below. Some of the data sets will be developed (particularly the ISA-Tab metadata and outputs derived from the raw data through data processing pipelines) in a Github repository.

The MetaboLights repository meets all requirements of the FAIR criteria. Each dataset will obtain an MTBLS accession number. MetaboLights studies can be claimed by the submitters on ORCID. While the accession number is persistent and unique, it is currently not equivalent to a DOI. In the future, DOIs might be registered by EMBL-EBI for MetaboLights studies.

Within the rules of MetaboLights, we'll use the naming conventions established in the creator's institution for experiments or samples.

2.2. Making data openly accessible

Experimental LC/MS metabolomics data

The MetaboLights repository is hosted at the EMBL-EBI and is a database for metabolomics experiments and derived information. MetaboLights accepts and stores all types of metabolomics data. According to EMBL-EBI terms of use, all public datasets are open and available for any purpose.

Data in MetaboLights is accessible through the web interface using the https protocol. Both raw data and metadata data is also available through the FTP[1], rsync[2] and Aspera[3] protocols. Programmatically the MetaboLights repository has a REST interface[4] to create, manage, search and access the studies.

Similar sections on GSMN

The project will make use of publicly available GSMNs. The C. elegans consensus GSMN WormJam is publically available on GitHub (https://github.com/JakeHattwell/wormjam). Currently the developmental version is available as SBML and SBTab files and will be followed up by regular releases. Any curation of network data will be handled using a GIT repository, allowing tracking of changes.

Human and Tomato GSMN will be retrieved from publications and associated repositories under SBML formats and then uploaded in MetExplore. Each metabolic network will be referred to with a unique MetExplore id. Network date of upload, reference will be kept in order to ensure reproducibility of future analysis. Any curation of network data will be handled using a GIT repository, allowing tracking changes.

Use and development of Ontologies

The project will make heavy use of existing chemical ontologies. The first one is ChEBI, developed at the EMBL-EBI. The ontology is documented and available at https://www.ebi.ac.uk/chebi/, and integrated into the Ontology Lookup Service https://www.ebi.ac.uk/ols/ontologies/chebi. Secondly, we are using ChemOnt developed in the Wishart lab. ChemOnt is interoperable with ChEBI, and accessible at and available for download from http://classyfire.wishartlab.com/. Work with upstream authors is underway to integrate ChemOnt into the OLS as well.

Open Source Software

All software developed in the scope of MetClassNet will be under an established Open Source License. Source code and documentation developed within MetClassNet will be hosted in GitHub repositories at https://github.com/MetClassNet

We will submit our contributions to Open Source software that needs modifications to work in the context of MetClassNet to the respective upstream project. Examples include

- MetExplore : https://forgemia.inra.fr/metexplore
- ISA-Tools https://github.com/ISA-tools/

[1] ftp://ftp.ebi.ac.uk/pub/databases/metabolights/studies/public/

[2] `rsync -rlpt -v -z rsync.ebi.ac.uk::pub/databases/metabolights/studies/public/MTBLS2 .`

[3] https://www.ibm.com/products/aspera

[4] https://www.ebi.ac.uk/metabolights/ws/api/spec.html#!/spec

Additionally, the following R and Bioconductor packages will be used and extended:

- XCMS https://github.com/sneumann/xcms/
- MSnbase https://github.com/lgatto/MSnbase
- rmzTab-M https://github.com/lifs-tools/rmzTab-m
- MetNet https://github.com/tnaake/MetNet

In some cases, we aim to replace existing in-house solutions in proprietary Matlab with Open Source alternatives.

2.3. Making data interoperable

LC/MS raw data will be available at least in the PSI standard format mzML[5]. For processed data we will use the PSI data standard mzTab-M[6] and the spectral library formats MSP or MGF for MS/MS spectra.

Identified metabolites will be reported with specific chemical structures, as far as they can be derived from the experimental conditions. Metabolites in GSMNs will be identified on their unique identifier in the model and the associated chemical structure. ChEBI will be used as a primary information source, but other databases will be used if the structure of interest has not (yet) been deposited to ChEBI.

The ISA-Tab metadata format makes heavy use of ontologies available through the Ontology Lookup System (OLS)[7] at EMBL-EBI.

2.4. Increase data re-use (through clarifying licences)

The MetaboLights repository is hosted at the EMBL-EBI and is a database for metabolomics experiments and derived information. MetaboLights accepts and stores all types of metabolomics data. According to EMBL-EBI terms of use, all public datasets are open and available for any purpose.

The MetaboLights submission pipeline is utilising the ISA software suite. All experimental data is extensively annotated in ISA-Tab format. MetaboLights enforces rigorous annotation requirements, set out in the MSI recommendations. Additionally, MetaboLights requirements for both raw and open-source data formats ensure that the primary research data is easily reusable.

According to MetaboLights guidelines, submitters are required to de-anonymise and pre-filter all datasets prior to submission to the archive. Submissions do not include consent forms, ethical approval or patient information and are governed by EMBL-EBI terms of use.

All data in the MetaboLights database is publicly available after curation approval and reaching the submitters' embargo date. In accordance with EMBL policy, the daily operation and running of this strategic archive is centrally funded and is therefore maintained without the need for short to medium-term funding.

[5] http://www.psidev.info/mzML
[6] https://github.com/HUPO-PSI/mzTab/#current-activities-and-software-support and https://hupo-psi.github.io/mzTab/2_0-metabolomics-release/mzTab_format_specification_2_0-M_release.html
[7] www.ebi.ac.uk/ols/

3. Allocation of resources

The data management activities in the project will be overseen by Reza Salek and Steffen Neumann, and executed by the respective data set owners.

All services and repositories used for developing and hosting the data sets are free of charge. The personnel costs to develop and submit are the responsibility of the owning institution, and they are readily covered through the ANR-DFG project grant. MetExplore and INRAE developments are hosted at the INRAE data centre. Costs are covered by the institute and by regional grant (CPER, MetaboHub-Metatoul).

4. Data security

Data submitted to **MetaboLights** will be under the standard data security processes at EMBL-EBI, which include geographically separated, redundant data centers.

At **HMGU,** generated LC-MS raw data is stored locally on the instrument PC and copied to a file server. This server has a double back-up in the IT department of the HMGU. Most of the data will be produced in vendor-specific file formats. Data is archived for at least 10 years. Processed data together with all metadata has an independent double back-up and is archived for at least 10 years.

At **IPB** various types of digital research data originate from the sample characterization by instrumental methods (here mainly LC-MS). Most of the data will be produced in vendor-specific proprietary file formats. Immediately after measurement, IPB policy mandates to store this ”primary research data” in IPB’s central primary data archive, which is where data is write-protected 7 days after its deposition (WORM functionality – Write Once Read Many) and archived for at least 10 years. Backup to tape is performed according to a defined schedule.

As a United Nations Agency, **IARC** falls under the overall umbrella of the United Nations Security System and must comply with relevant regulations. IARC has three storage bays providing over 200TB of storage for Scientific activities including High Performance Computing (HPC) and nearly 100TB of storage for our central servers, data storage and backups, used for storing metabolomics data. The IARC network is a highly redundant architecture with multiple links from different sections with the institute. IARC HPC is a private cloud running on OpenStack to provide on-demand access to differently configured virtual machines. All the raw and processed data are stored locally on a portable disk as well as on the remote servers. We keep at least 3 different copies for each study.

At **INRAE**, data are stored in servers hosted by Genotoul-bioinformatics facility (Toulouse) membre of French Bioinformatics Institute (IFB). Virtual Machines (MetExplore various servers) are also hosted at this facility.

As metabolomics data available at IARC, is human data, publicly sharing this data is not possible at this time. However, we are working on a solution to make this data available, while keeping some of the

5. Ethical aspects

Data obtained at IARC are mainly from human clinical studies. The human clinical data has been obtained under local ethical and legal requirements and stored locally in an anonymised form in IARC (see below). Sharing these data requires approval and preparation of a Material Transfer

Agreement (MTA). As IARC, IARC Biobank (IBB) is a centralized biological resource storage facility for samples collected from studies conducted worldwide by IARC in collaboration with international partners (http://ibb.iarc.fr/). The Laboratory Services and Biobank Group (LSB) is responsible for the management of the IBB. The Group also provides services in pre-analytical sample processing and shipment. IARC Ethics Committee (IEC) is to ensure the protection of the safety, well-being and privacy of participants in studies carried out by IARC. The IEC reviews, approves and monitors all research projects coordinated by IARC scientists or in which they participate. All projects need to go through to IEC clearance procedures. In particular, ethical approval from the IEC is required for all research proposals involving individuals recruited especially for the study, collections of human biological materials, or data allowing individuals to be identified, in any study in which IARC is involved in any capacity (see also WHO Manual XV.3.1 paragraph 1002). For more details and procedures see http://ethics.iarc.fr

Data obtained at HMGU will be from *C. elegans* and not involve any human subjects. Data obtained at IPB will be from plants and not involve any human subjects.