Gaining a structural perspective on Ne-lysine acetylation in bacteria and their bacteriophages

A Data Management Plan created using DMPonline.be

Creator: Hannelore Longin

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 11PCC24N

ID: 206389

Start date: 01-11-2023

End date: 31-10-2027

Project abstract:

Bacterial viruses, or phages, have played a pivotal role in deducing the foundations of molecular biology and inspired indispensable biotechnological tools. During phage infection phages 'hijack' the bacterial metabolism in efficient and diverse manners. We recently discovered a novel layer of phage-based regulation: acetylation-based post-translational modifications of proteins in infected cells. While lysine acetylation has proven to be a crucial post-translational modification in eukaryotes, functional insight is lacking in prokaryotes. To date, lysine acetylation studies in prokaryotes have mainly focused on large scale identification of target proteins. However, we believe recent advances in structural bioinformatics (AlphaFold2) present a unique opportunity to study the biochemical impact of lysine acetylation at the level of individual proteins on a proteome-wide scale in prokaryotes.

In this proposal, we will model all known lysine acetylations in*Pseudomonas aeruginosa* and predict whether they (de)stabilize protein interactions, by applying a self-developed bioinformatics framework. In addition, we will identify and characterize phage-encoded acetyltransferases. Finally, focused case studies will allow more biochemical details to be revealed. This will lead to functional insight into bacterial lysine acetylation and how phages impact this, which could inspire biotechnological advances.

Last modified: 17-04-2024

Gaining a structural perspective on Ne-lysine acetylation in bacteria and their bacteriophages
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	
Pseudomonas acetylomics data	Existing in-house and published acetylomics data for *Pseudomonas aeruginosa*	Reuse existing data	Digital	Experimental	.xlsx	<100 MB	
Protein interaction data	*Pseudomonas* protein interaction data from STRING	Reuse existing data	Digital	Compiled/aggregated data	.txt	<100 MB	
Protein annotation data	*Pseudomonas* (phage) protein sequence (annotations)	Reuse existing data	Digital	Experimental	.fa	<100 MB	
Acetyltransferase HMMs	Pre-made Hidden Markov Models for acetyltransferases	Reuse existing data	Digital	Experimental	.hmm	<100 MB	
Pre-exisiting protein structure models	Pre-calculated /experimentally determined protein structures	Reuse existing data	Digital	Experimental; Simulation data	.pdb; .mmcif; .json	<100 GB	
Own protein structure models	Own AlphaFold models of *Pseudomonas* protein (interactions); phage proteins	Generate new data	Digital	Simulation data	.pdb; .json; .png; .txt; .log	<100 GB	
Protein properties	Calculated protein (pocket) properties	Generate new data	Digital	Aggregated data	.txt	<1 GB	

Molecular dynamics simulations & analysis	Data related to molecular dynamics simulations and MM-GBSA calculations	Generate new data	Digital	Simulation data	.crd; .pdb; .top; .gro; .xtc; .ndx; .dat	<5 TB	
Bacterial two-hybrid data	Data related to bacterial two hybrid experiments	Generate new data	Digital	Experimental	.jpg; .tiff; .png; .xlsx; .csv	<1 GB	
Western Blot data	Data from Western Blot experiments	Generate new data	Digital	Experimental	.jpg; .tiff; .png	<1 GB	
Phage (protein) acetylomics data	Acetylomics data related to phage acetyltransferases	Generate new data	Digital	Experimental	.xlsx	<100 MB	
Phage infection data	Data related to phage infection experiments	Generate new data	Digital	Experimental	jpg; .tiff; .png; .xlsx; .csv	<100 MB	
Biological data	Bacterial, viral strains and various DNA constructs	Generate new data	Physical	Experimental	NA		part of shelf in lab freezer
Code	Pipelines/scripts written during the project	Generate new data	Digital	Software	.py; .sh; .ipynb	<1 GB	

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:

- *Pseudomonas* acetylomics data:
  - PAO1 data deposited in proteomics database ProteomeXchange:
    - ProteomeXchange identifier: PXD051454
  - PA14 data extracted from publications:
    - Gaviard publication:
      - doi: 10.1021/acs.jproteome.8b00210
      - data from supplementary table S-3
    - Ouidir publication:
      - doi: 10.1002/pmic.201500056
      - data from supplementary table S-1
- Protein interaction data:
  - data extracted from STRING-db (https://stringdb-downloads.org/download/):
    - PAO1 data:
      - TAX ID: 287
      - STRING version: v11.5
      - URL: https://stringdb-downloads.org/download/287.protein.physical.links.full.v11.5.txt.gz
    - PA14 data:
      - TAX ID: 208964
      - STRING version: v11.5
      - URL: https://stringdb-downloads.org/download/208964.protein.physical.links.full.v11.5.txt.gz
- Protein annotation data:
  - Phage protein data extracted from NCBI protein:
    - List of NCBI protein identifiers (tens of thousands) will be stored in a dedicated file, stating date of retrieval
    - URL: https://www.ncbi.nlm.nih.gov/protein/
  - *Pseudomonas* protein data extracted from UniProt:
    - List of UniProt protein identifiers (thousands) will be stored in a dedicated file, stating UniProt release on the date of retrieval
    - URL: https://www.uniprot.org/
- Acetyltransferase HMMs:
  - identifiers and repositories will be added once HMMs have been selected
- Pre-exisiting protein structure models:
  - Protein structure models extacted from PDB:
    - List of PDB model identifiers will be stored in a dedicated file

        - URL: https://www.rcsb.org/
    - Protein structure models extracted from AlphaFold database
        - List of AlphaFold model identifiers will be stored in a dedicated file
        - URL: https://alphafold.ebi.ac.uk/

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

Yes, this project can result in research data with potential for tech transfer and valorisation. If applicable, potential IP restrictions for obtained data will be evaluated with KU Leuven Research & Development (LRD). When the IP is patented or the procedure has been stopped, the research data will be published in peer-reviewed articles.
It is currently hard to predict which part/data of the project will have commercial potential. This section will be updated once commercial valorization becomes more concrete.

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.

- No

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.

- No

All reused data is publicly available open access data; and all generated data will be generated in-house or through commercial partners, maintaining full access and rights to the lab.

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).

All experimental, generated data and how it was processed is tracked using digital journals/lab books, which are kept on university-secured network drives. Both lab books and data are time-stamped, allowing to easily trace back (experimental) details of the corresponding data.

Where possible we will also strive to include how data was processed into the file or folder of the respective processed data.
All other (computational) data will be documented by use of READMEs and code Notebooks. Python code will be documented with the Google docstrings format (in the code files).

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

Metadata will follow the standards of the relevant repositories (will be added in due time). Data published in or as supplement of open-access peer-reviewed publications will conform to the standards of the publisher.

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

Where will the data be stored?

Code will be stored in KU Leuven Gitlab repositories. All datasets, documentation and metadata will be stored on the research group L-drive within the secure KU Leuven environment. Access to this drive is limited to the PI and involved researchers.
Code and datasets used locally will be maximally stored in the KU Leuven Onedrive (provided for each staff member). Any changes or updates will be uploaded weekly to the L-drive and KU Leuven Gitlab.
In case data is linked to publications, it will be also made available on public databases which have their own storage facilities.
Biological data, comprising bacterial and viral strains and DNA constructs, are stored in freezers (-20°C and -80°C) present in the laboratory. The samples are labelled and time-stamped, allowing easily track back in lab books and in databases of the corresponding freezers.

How will the data be backed up?

KU Leuven provides automatic backup for their L-drives. Automatic version management occurs when storing data in the KU Leuven datacenters. "Snapshot" technology is used for version management, which keeps the previous versions of the changed file online on the same storage system. Additionally, KU Leuven mirrors these files to a second ICTS data center to enabledisaster recovery in case of problems. Files are versioned once a day, and version are kept for 14 days. KU Leuven Gitlab ensures automatic backups, and all file versions are permanently stored.
For important biological data, a back-up sample will be made and stored in a second -80°C freezer present in the laboratory.

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

Our research group currently has an L-drive with a capacity of 5 TB for research. This should be sufficient for storage of the generated datasets. In case of capacity problems, the L-drive capacity can be expanded or data can be stored on the KU Leuven OneDrive for staff (4 TB).
Sufficient storage capacity for physical samples in both -20°C and -80°C freezers, as well as temporary storage at 4°C, is available in the laboratory.

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

The storage facilities of KU Leuven ensure secure storage of data, with restricted access control. Only specific lab members will have access to the shared folder, unauthorized persons do not have access to this system.
A badge system allows only access to the laboratory for authorized people.

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

The L-drive storage has a yearly cost of 104,42/TB/year and is covered by the research group. Storage for more than 5 years after the project can be covered by the research group/department.

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...).

All digital data, that is not already deposited in a public repository, will be preserved and will remain available for members of the labs involved in this project.
Important biological samples (e.g. engineered bacterial and viral strains, specific plasmids) will be stored in -20°C/-80°C freezers available in the laboratory. Other biological materials (e.g. primers, intermediates, unfinished products) will be discarded during and at the end of the project due to limited storage capacity in the long-term storage freezers.

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

Digital data will be stored on the university's central servers (with automatic back-up procedures) for at least 10 years, conform the KU Leuven RDM policy.
Specific long-term storage freezers (-20°C and -80°C) are available in the laboratory. Biological sample details will be listed in databases, which are stored on the university's central server.

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

No extra costs are expected for data preservation.

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

Results coming from the project will be published and associated datasets will be either included in supplementary materials of the publication or be made available through public repositories. This is expected to apply to all described datasets.
The data will also be stored and be available for lab members using a shared network drive provided by the KU Leuven.

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

All members of the labs involved in the project will be authorized to have access to the obtained data. Data which is not under specific IP, confidentiality or other agreements will be shared in public databases after publication whenever possible, if not, this data can be shared upon request by academic researchers.

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.

- Yes, Intellectual Property Rights

Yes, patents and confidentiality agreements are the main factors that can limit data sharing.

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

Results and datasets from this project will be shared in common public repositories, unless it is customary to only provide the data type as supplementary material. An overview of the expected repositiory per data type is given below:

- Own protein structure models: Zenodo
- Protein properties: Zenodo
- Molecular dynamics simulations & analysis: Zenodo
- Bacterial two-hybrid data: as supplementary material to publication
- Western blot data: as supplementary material to publication
- Phage (protein) acetylomics data: ProteomeXchange
- Phage infection data: as supplementary material to publication
- Biological data: available upon request
- Code: GitHub

When will the data be made available?

Data will be made available upon publication and/or project end.

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

- Code: MIT license
- Datasets: CC-BY(-NC) 4 (depending on underlying data 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

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

There are no expected costs for data sharing, as the expected data sharing platforms (such as Zenodo, ProteomeXchange, and GitLab/GitHub) are free of charge.

6. Responsibilities

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

The PhD researcher (Hannelore Longin) will be responsible for data documentation and metadata during this research project.

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

The PhD researcher (Hannelore Longin) will be responsible for data storage and back up during this research project.

Who will manage data preservation and sharing?

The PhD researcher (Hannelore Longin) will be responsible for data preservation and sharing. The PIs (prof. Vera van Noort; Prof. Rob Lavigne) will be responsible for data preservation and sharing past this projects duration.

Who will update and implement this DMP?

The PhD researcher (Hannelore Longin) will be responsible for updating and implementing this DMP. The PI (prof. Vera van Noort) will monitor the DMP implemetation.