DMP title

Project Name DMP_G0C0422N - DMP title

Grant Title G0C0422N

Principal Investigator / Researcher Dirk Springael

Description Despite the unnatural chemical structure of pesticides and the short time frame in which they occur in the environment, bacteria emerged that carry catabolic pathways for using the chemicals for growth. Although originating from distant geographical areas, isolates that degrade the same pesticide, display basically similar catabolic pathways and genes, although with differences in gene sequence and organization. The questions are whether such genotypes emerged independently and/or whether they spread and further evolved from a few places and how their genetic make-up and evolution contributed to their distribution. Due to the limited number of isolates from a limited number of locations and the fact that only information is available from cultured strains, these questions can currently not be answered. This project will do so by interrogating the worldwide variation of the tfd genotype that encodes for the catabolism of the herbicide 2,4-dichlorophenoacetic acid (2,4-D) using culture-dependent and culture independent approaches. It will be examined whether region-specific 2,4-D catabolic tfd genotypes and profiles of global distribution and evolution can be recognized. This will be linked to biotic and abiotic soil parameters and climatic and anthropogenic parameters to infer the underlying factors of emergence/ evolution/distribution of the tfd genotypes and the role of the tfd mak-up in global spread.

Institution KU Leuven

1. General Information

Name applicant

Dirk Springael

FWO Project Number & Title

G0C0422N: World-wide diversity of genotypes determining the bacterial catabolism of a pesticide: does it tell us something about the global origin, evolution and distribution of xenobiotic catabolic gene functions?

Affiliation

- KU Leuven

2. Data description

Will you generate/collect new data and/or make use of existing data?

- Generate new data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Metadata files (.xls) on origin and characteristics of samples, around 30 in total (max l MB) created by gathering information on site and physico-chemical analysis of soil characteristics.

Sequencing data files (fastq format), around 200 (plasmid, metagenome sequences), approx. 6 TB, generated by short read/long read sequencing + 16S amplicon sequencing data files (fastq format) from around 60 from soil samples generated by Illumina sequencing, approx. 100 GB.

OTU table derived from the 16S amplicon sequences and containing the relative abundances of each Operational Taxonomic Unit (OTU) for each sample (in.csv format)

Bacterial culture collection (approx. 500 strains) in glycerol medium, stored at -80°C. 2 replicates per strain.

Mineralization kinetic data (.xls), around 60 in total (max. 10 MB) generated by lab-experiments collecting $^{14}CO_2$ produced from 14C-labeled 2,4-D

3. Legal and ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for

Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- No

Privacy Registry Reference:
Short description of the kind of personal data that will be used:

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s)

- No

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- Yes

We have to check this out. We will use soil samples from different countries all over the world and look for the presence of specific bacteria and genes. So, our research, will involve genetic resources that might be captured by the EU Regulation related to the Nagoya Protocol. We will though never really "use" these genetic resources for applocations etc.

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

- No

4. Documentation and metadata

What documentation will be provided to enable reuse of the data collected/generated in this project?

Overall, (e-)labbooks will contain information on experimental design, protocols, sampling location, abbreviations used, structure of the data (including link with physical storage of data), and steps involved in data analysis and relevant analysis scripts (R scripts, MOTHUR/QIIME scripts). A clear coding for all data files related to the project will be used. In the concluding stage of the project, a master index file containing the combined information for all experiments will be compiled which will be archived and also stored on the personal harddrives/PC of the PI. Altogether, this should allow any secondary analyst to use the data accurately and effectively.

More specifically, the following information will be given on the items described in section 2:

Metadata files on origin and characteristics of samples, will be provided with a clear description of the methods (like soil analytics) that were used to collect the data. The metadata will include information on the samples (where collected (GPS coordinates), environmental conditions at site when isolated, history of the location, soil physico-chemical characteristics). In addition it will provide where the related sequence and mineralization data can be found.

Sequencing data files deposited in sequence data bases like EMBL will include the information/ documentation required by the data base.

OTU tables derived from 16S amplicon sequences and containing the relative abundances of each Operational Taxonomic Unit (OTU) for each sample will be deposited in an official nucleic acid database like EMBL and contain the documentation requested by the depository.

The bacterial cultures isolated in the study will be preserved as freezer stocks in glycerol in our laboratory collection at -80°C, and a file with strain details (in Access) (identification/surce of origin/main characteristics/storage medium/revival guide/location in the freezer) will be maintained. We have a dedicated on line own system for that.

Mineralization kinetic data will be provided including a detailed description of the method for collecting the data.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- No

Metadata files on origin and characteristics of samples, will be provided with a clear description of the methods (like soil analytics) that were used to collect the data. The metadata will include information on the samples (where collected (GPS coordinates), environmental conditions at site when isolated, history of the location, soil physico-chemical characteristics). In addition it will provide where the related sequence and mineralization data can be found.

Sequencing data files deposited in sequence data bases like EMBL will include the information/ documentation required by the data base.

5. Data storage and backup during the FWO project

Where will the data be stored?

All data will be stored in the university's secure environment. Nucleic acid sequence data (plasmid, metagenome, 16S amplicon) will be submitted and stored in official nucleic acid data bases like EMBL upon publication.

How is backup of the data provided?

The data will be stored on the university's central servers with automatic daily back-up procedures.

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

The amount of data foreseen to not encompass foreseen capacities

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

If any, not expected to be high. Costs will be covered by the FWO project itself.

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The data will be stored in the university's secure environment.

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

In the concluding stage of the project, a master index file containing the combined information for all experiments will be compiled which will be archived and also stored on the personal harddrives/PC of the PI. Altogether, this should allow any secondary analyst to use the data accurately and effectively.All data will be preserved for at least 5 years after completion of the project.

Where will the data be archived (= stored for the longer term)?

The 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. Nucleic acid sequence data will be stored at official public data bases like EMBL.

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

No direct idea but not expected to be high. If costs, they will be covered by related research projects.

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

- No

Which data will be made available after the end of the project?

The full dataset (except the nucleic acid sequence data deposited at official repositories for sequencing results) will be deposited in a cvs format in KU Leuven RDR under a CC-BY license. The nucleic acid sequence data will be avalaible through official nucleic acid databases like EMBL.

Where/how will the data be made available for reuse?

- In an Open Access repository

Through KU Leuven RDR.

When will the data be made available?

- Upon publication of the research results

The full dataset will be uploaded and made available in a cvs format in RDR immediately afer the end of the project in case published. Others will be added upon publication. If not published within 1,5 years of project complation, all datasets will be made available.

Who will be able to access the data and under what conditions?

Open access data in RDR.

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

No high costs expected. In case, costs will be covered by FWO project itself.

8. Responsibilities

Who will be responsible for data documentation & metadata?

Anahita Modabberi who will act as the PhD student working on the project.

Who will be responsible for data storage & back up during the project?

Anahita Modabberi who will act as the PhD student working on the project.

Who will be responsible for ensuring data preservation and reuse ?

PI Dirk Springael

Who bears the end responsibility for updating & implementing this DMP?

The PI bears the end responsibility of updating & implementing this DMP.