MICROBIOME

WG3 progress reporting & future plan

25th October 2021

Magali Berland
Michelangelo Ceci
Sonia Tarazona

WG3 Objectives and major deliverables

Objectives:

To optimise and standardize the use of state-of-the-art ML techniques, resulting in **best practice SOPs** specific to various microbiome data types, human body ecosystems and research questions. The WG3 will also investigate opportunities for **automating** the established SOPs into pipelines for translational use by clinicians and non-experts.

Major Deliverables:

D3.1: [oct - dec 2021] A decision tree of ML/Stats methods along with optimised parameters suitable for various data types, ecosystems and research questions (disseminated through Web-portal and GitHub).

D3.2: [april - jun 2022] A publication and white-paper describing the SOPs emanating from D3.1.

D3.3: [july - sept 2022] A report outlining areas suitable for automation

Summary of WG3 progress (2019-2021)

Several threads of research, from different groups and collaborations. For the moment, mainly analysis on publicly available datasets (mainly 16s). (**...this is not an exhaustive list**)

- Explain the observed diversity in human microbiome (*University of Turku, Finland*)
- Predicting the onset of Type2 diabetes with AutoML using microbiome data *(Dept. of Computer Science, University of Bari Aldo Moro, Bari, Italy - Institute of Genomics, University of Tartu, Tartu, Estonia*)
- Probabilistic distribution of taxonomic units (*Ss. Cyril and Methodius University in Skopje, North Macedonia*)
- Clustering and classification of human microbiome data (*University of Novi Sad, Serbia- University of Ljubljana, Slovenia*)
- Comparing different normalization strategies and ML methods on 6 different datasets for 5 diseases (*Universitat Politècnica de València, Spain*)
- Analysis of human microbiome data with JADBio (*Department of Computer Science, University of Crete, Greece, FORTH*)
- Statistical and ML analysis of microbiome data using the logratio methodology of compositional data (*Palacký University, Czech Republic)*

Short talks (07/07/21)

- Karel Hron: ***Why are microbiome data compositional?***
- Andrea Mihajlovic (Tatjana Loncar Turukalo): ***Inflammatory bowel disease prediction based on metagenomics data***
- Magali Berland: ***Extensive benchmark of machine learning methods for microbiome data***
- Michelangelo Ceci: ***Predicting the onset of Type2 diabetes through the analysis of microbiome data***

Summary of WG3 progress (2019-2021)

From the studies of the members we started to define a decision tree for SOP, showing in different data/normalization/pre-processing/algorithms what is the best approach according to their experience

Example:

DECISION	OPTIONS		
Data type?	Shotgun Metagenomics		
Pre-processing pipeline?	?????		
Variable filter?	None	Low counts	
Normalization/Transformation?	TSS	CLR	
Type of method?	Unsupervised	Classification	Regression
Algorithm?	RF	NN	SVM

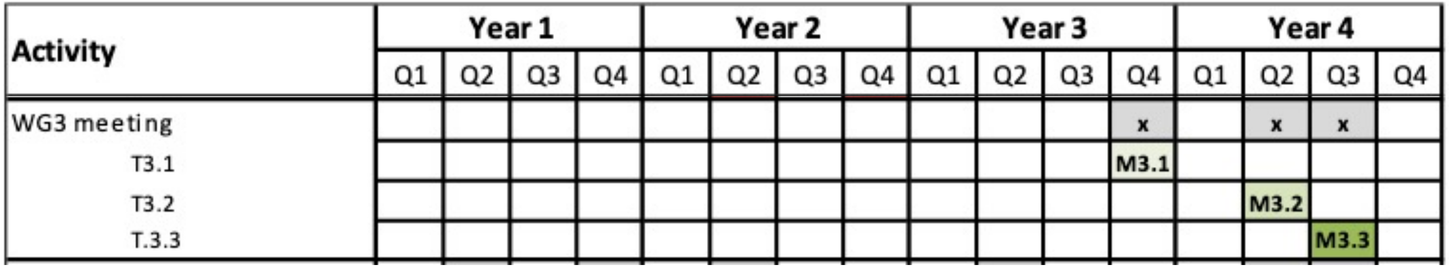
Summary of WG3 progress (2019-2021)

Two main approaches to choose the Operating Procedures to be adopted in the studies:

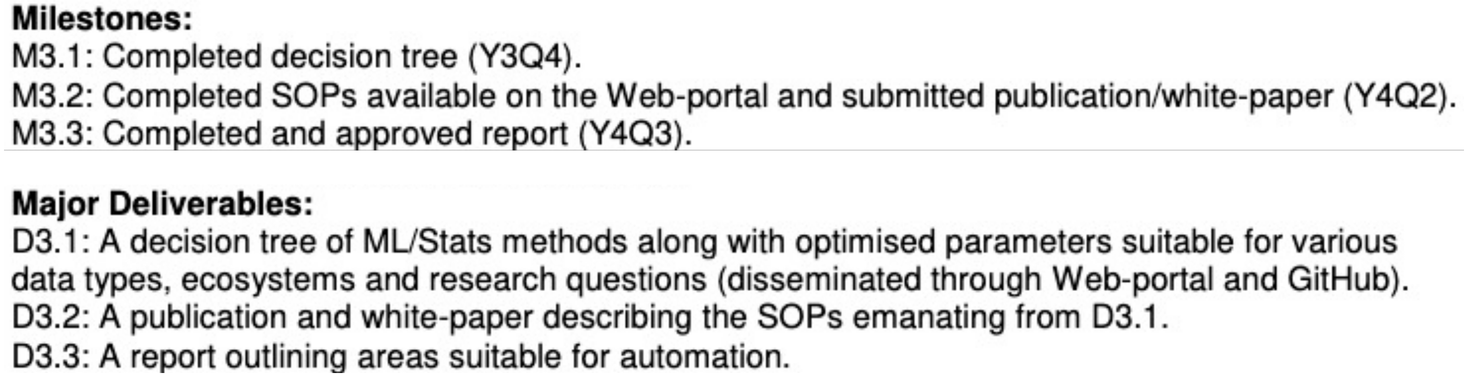
- Classical experimental & explanatory approach
- Automatic, based on AutoML

A joint work with WG1 has also been conducted to identify and analyze relevant papers. The standards steps from existing literature will also be included in the tree when relevant.

Progress and deliverables on the Gantt Chart



Activity	Year 1				Year 2				Year 3				Year 4			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
WG3 meeting												x		x	x	
T3.1												M3.1				
T3.2														M3.2		
T.3.3															M3.3	



Milestones:
M3.1: Completed decision tree (Y3Q4).
M3.2: Completed SOPs available on the Web-portal and submitted publication/white-paper (Y4Q2).
M3.3: Completed and approved report (Y4Q3).

Major Deliverables:
D3.1: A decision tree of ML/Stats methods along with optimised parameters suitable for various data types, ecosystems and research questions (disseminated through Web-portal and GitHub).
D3.2: A publication and white-paper describing the SOPs emanating from D3.1.
D3.3: A report outlining areas suitable for automation.

Monthly meetings

- July 7th – 12 Participants
- September 10th - 10 Participants + 1 cannot be present
- October 19th - 13 Participants + 2 cannot be present
- November 17th

Main discussion topics

- The dataset saga:
  - shotgun = Microbiome atlas, then CRC cohorts
  - 16S = ???
- Variable filter / normalization / transformation
- Pipelines for the optimization and standardization step (~ 5-7 teams on the task)
- Decision Tree building from literature – need help from WG1
- Compositional data analysis

Support for other WGs

- WG1 State-of-the-art evaluation and update
- **WG2 Benchmark data & DREAM Challenge**
  - Input datasets for optimisation & standardisation tasks
- WG3 Optimisation and standardisation
- **WG4 Dissemination and training**
  - October ML4microbiome training school – several trainers
  - ML4Microbiome workshop – 2 lectures (EMBnet & GOBLET Annual General Meeting 2021)