COST
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

ML4 MICROBIOME

ML4MICROBIOME WORK PACKAGE 1

State-of-the-art evaluation and update

Overview of activities

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

30.08.2022

Objectives

To continuously evaluate the state-of-the-art ML/statistics methods, and to ensure that every action member is "on the same page" in terms of their robustness and suitability for microbiome research and how well they address the specific challenges in 1.1.1, separately and combined.

Task 1.1: Technology watch

Task 1.2: Evaluation of ML/statistics methods currently used in microbiome research

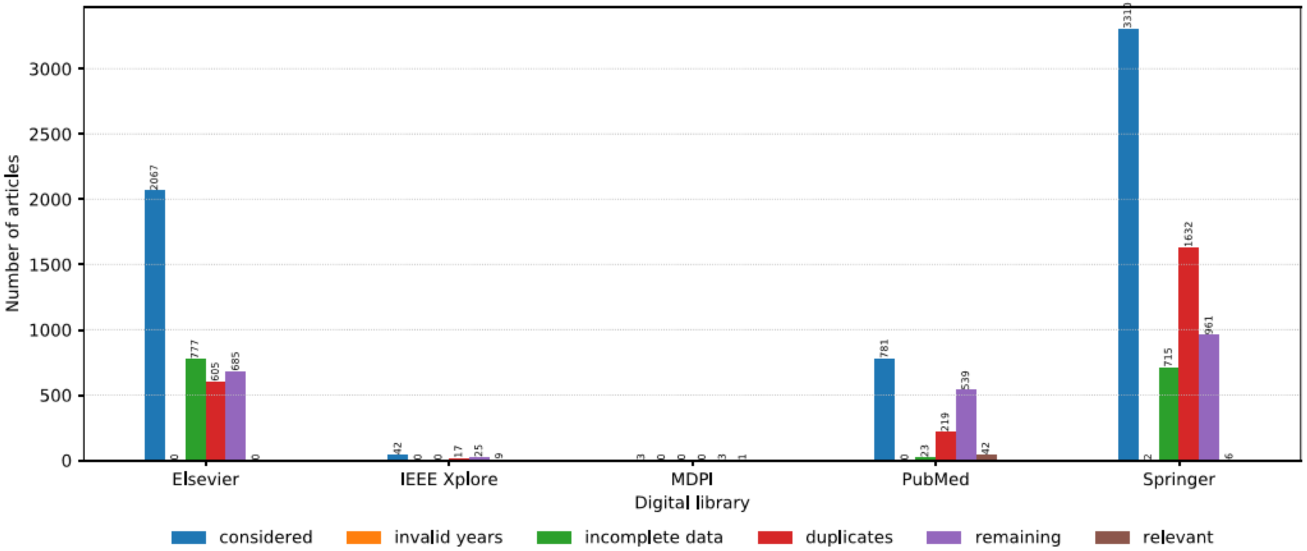
Task 1.3: Define priority areas for novel ML/statistics applications for microbiome data

WG1 activities in 2021/2022

- Preparation of updated annual report for 2021
  - *Will be based on revised first annual report and review paper published in Frontiers in Microbiology*
  - *Compilation of the report will follow the same process as was used in previous year*
- Additional task
  - *Next action/WG1 publication – topic was discussed and fixed among WG1 members. Writing plan was fixed at Tirana meeting.*

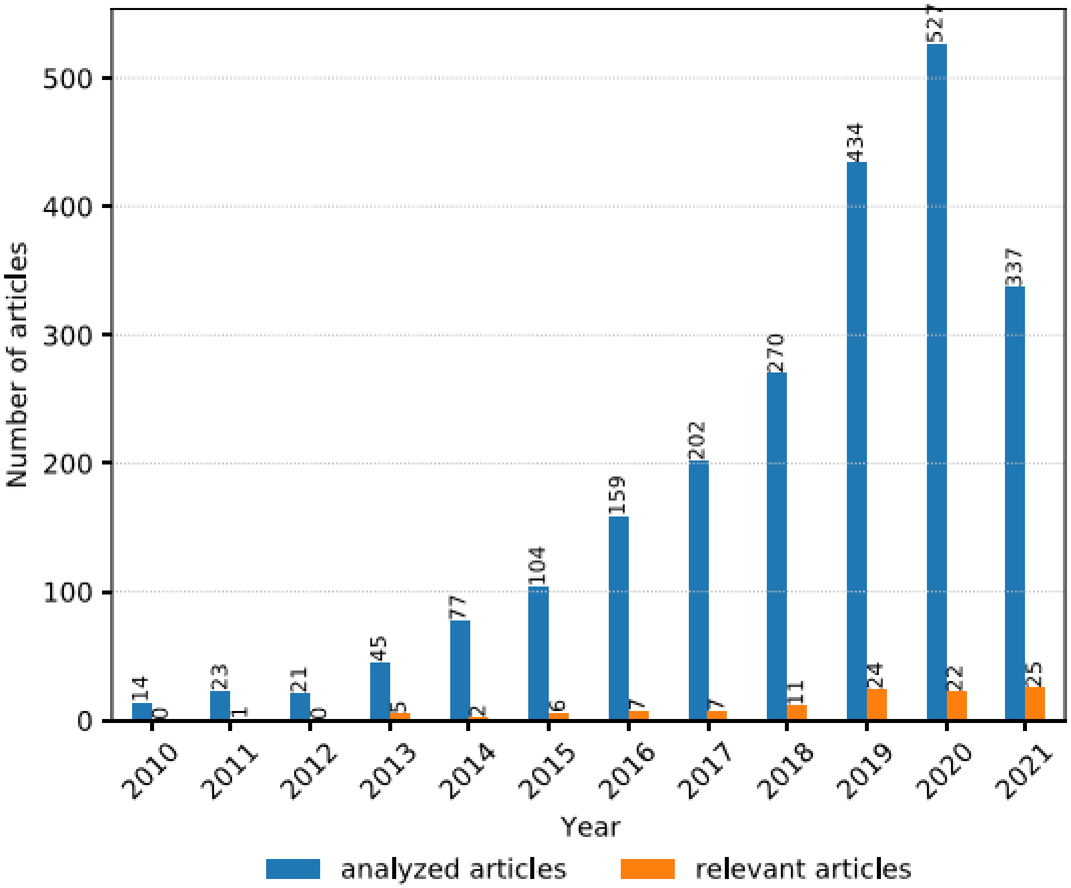
Scoping Review

- In 2020, Springer, IEEE, and PubMed were accessed
- In 2021, two additional digital libraries: Elsevier and MDPI were included
- No Oxford Academic Publishers included



NLP toolkit (Zdravevski et al., 2019)

Publication search results



Microbiome research in the source code repositories

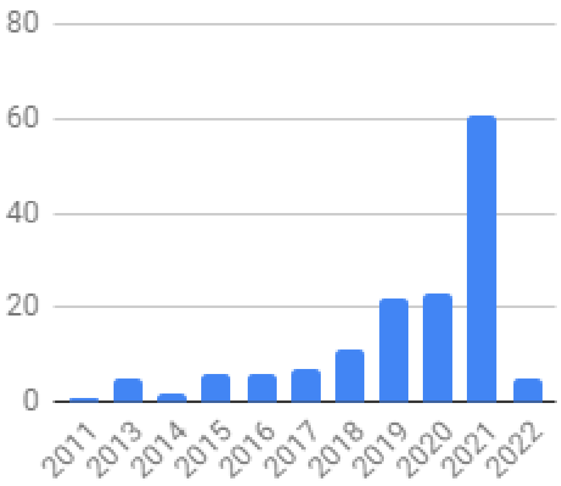
Short summary (from September 2021)

- Number of repositories 1855 (old), 2465 (new).
- New repositories 1014.
- Repositories deleted 404.

- Update (October)
  - As of today we have slightly more repositories (2589).
- Updated info on github list http://microbiome.przymus.org/
- Next step – identify new (unique) publications for 2021
- Number publications for year 2021 – 4 (Dec 2021)

Performed tasks

1. Data base curation – Scoping review
2. Data base creation/curation – Github
3. Additional search (OAP)
4. Merging the obtained data bases
5. Updating the data base by action members (Feb-March 2022)
6. Report compilation (June-Aug 2022)



Article data base generation

- An automated search of digital libraries of three major publishers (PubMed, Springer, Elsevier, MDPI and IEEE) using NLP Toolkit (Zdravevski et al., 2019) to automate the literature search, scanning, and eligibility assessment. This method yielded **25 papers**.

- An automated search through the available GitHub resources using NLP algorithms to identify relevant software repositories and extract corresponding scientific papers. The papers were automatically ranked by relevance using the pointwise learning to rank approach (Fejzer et al. unpublished) trained using the manually collected and labeled papers. The final list includes **four papers**.

- Manual search – crowdsourcing of the studies relevant for the review topic by all members of the COST Action CA18131 “Statistical and machine learning techniques in human microbiome studies”. In this way, **33 papers** were added to the final list.

After revision - **41papers** in final list.

Analysis of collected article data set

- 41 papers were adopted for the year 2021, indicating a substantial increase in the application of ML methods for human microbiome analysis compared to previous years.
- The primary disease type in collected articles was inflammatory bowel disease (19%), followed by drug-related side effects and adverse reactions, and diabetes.
- More than 70% of studies have used amplicon sequencing data (16S rDNA) and 10% only shotgun metagenome data as input data type.
- The most often used methods were random forest, logistic regression and support vector machine. Comparison with an earlier data set shows that random forest is still the most used method, but the application of logistic regression and support vector machine algorithms has increased.

Analysis of collected article data set - reviews

- Curry, K. D., Nute, M. G., & Treangen, T. J. (2021). It takes guts to learn: machine learning techniques for disease detection from the gut microbiome. Emerging Topics in Life Sciences, 5(6), 815–827. https://doi.org/10.1042/ETLS20210213
- Ghannam, R. B., & Techtmann, S. M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Computational and Structural Biotechnology Journal, 19, 1092–1107. https://doi.org/10.1016/J.CSBJ.2021.01.028
- Hajeebu, S., Ngembus, N. J., Bandi, P. S., Panigrahy, P. K., & Heindl, S. (2021). Machine Learning as a Tool in Investigating the Possible Role of Microbiome in Development and Treatment of Cancer. Cureus, 13(8). https://doi.org/10.7759/CUREUS.17415
- McCoubrey, L. E., Elbadawi, M., Orlu, M., Gaisford, S., & Basit, A. W. (2021). Harnessing machine learning for development of microbiome therapeutics. Gut Microbes, 13(1), 1–20. https://doi.org/10.1080/19490976.2021.1872323
- McCoubrey, L. E., Gaisford, S., Orlu, M., & Basit, A. W. (2022). Predicting drug-microbiome interactions with machine learning. Biotechnology Advances, 54. https://doi.org/10.1016/J.BIOTECHADV.2021.107797
- Prihoda, D., Maritz, J. M., Klempir, O., Dzamba, D., Woelk, C. H., Hazuda, D. J., Bitton, D. A., & Hannigan, G. D. (2021). The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability. Natural Product Reports, 38(6), 1100–1108. https://doi.org/10.1039/D0NP00055H
- Wu, S., Chen, Y., Li, Z., Li, J., Zhao, F., & Su, X. (2021). Towards multi-label classification: Next step of machine learning for microbiome research. Computational and Structural Biotechnology Journal, 19, 2742–2749. https://doi.org/10.1016/J.CSBJ.2021.04.054
- Zhang, W., Chen, X., & Wong, K. C. (2021). Noninvasive early diagnosis of intestinal diseases based on artificial intelligence in genomics and microbiome. Journal of Gastroenterology and Hepatology, 36(4), 823–831. https://doi.org/10.1111/JGH.15500

Analysis of collected article data set – by action members

- Aasmets, O., Lüll, K., Lang, J. M., Pan, C., Kuusisto, J., Fischer, K., Laakso, M., Lusis, A. J., & Org, E. (2021). Machine Learning Reveals Time-Varying Microbial Predictors with Complex Effects on Glucose Regulation. MSystems, 6(1). https://doi.org/10.1128/MSYSTEMS.01191-20/SUPPL_FILE/MSYSTEMS.01191-20-ST003.DOCX
- Bakir-Gungor, B., Bulut, O., Jabeer, A., Nalbantoglu, O. U., & Yousef, M. (2021). Discovering Potential Taxonomic Biomarkers of Type 2 Diabetes From Human Gut Microbiota via Different Feature Selection Methods. Frontiers in Microbiology, 12, 1330. https://doi.org/10.3389/FMICB.2021.628426/BIBTEX
- Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., Berland, M., Gruca, A., Hasic, J., Hron, K., Klammsteiner, T., Kolev, M., Lahti, L., Lopes, M. B., Moreno, V., Naskinova, I., Org, E., Paciência, I., Papoutsoglou, G., … Truu, J. (2021). Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. Frontiers in Microbiology, 12, 313. https://doi.org/10.3389/fmicb.2021.634511
- Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., Aydemir, O., Bakir-Gungor, B., Santa Pau, E. C. de, D’Elia, D., Desai, M. S., Falquet, L., Gundogdu, A., Hron, K., Klammsteiner, T., Lopes, M. B., Marcos-Zambrano, L. J., Marques, C., Mason, M., … Claesson, M. J. (2021). Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. Frontiers in Microbiology, 12, 277. https://doi.org/10.3389/fmicb.2021.635781
- Patel, S., Vlasblom, A. A., Verstappen, K. M., Zomer, A. L., Fluit, A. C., Rogers, M. R. C., Wagenaar, J. A., Claesson, M. J., & Duim, B. (2021). Differential Analysis of Longitudinal Methicillin-Resistant Staphylococcus aureus Colonization in Relation to Microbial Shifts in the Nasal Microbiome of Neonatal Piglets. MSystems, 6(4). https://doi.org/10.1128/MSYSTEMS.00152-21/
- Theelen, M. J. P., Luiken, R. E. C., Wagenaar, J. A., Sloet van Oldruitenborgh-Oosterbaan, M. M., Rossen, J. W. A., & Zomer, A. L. (2021). The Equine Faecal Microbiota of Healthy Horses and Ponies in The Netherlands: Impact of Host and Environmental Factors. Animals 2021, Vol. 11, Page 1762, 11(6), 1762. https://doi.org/10.3390/ANI11061762

Other activities

- Next action/WG1 publication –

Manuscript draft about ML software for human microbiome analysis

- Presentations
  - *Enrique Carrillo - Machine Learning & Microbiome for Precision Nutrition. GOBLET & EMBnet AGM 202*

WG1 meetings in Turku

- Topics to discuss
  - Finalizing updated annual report for 2021
  - Next action/WG1 publication – Manuscript draft about ML software for human microbiome analysis
  - Updated annual report (D1.5) for the year 2022.
  - Report outlining priority areas for new ML/statistics methods (D1.7)

ML4MICROBIOME WORK PACKAGE 1

State-of-the-art evaluation and update

Summary of WG1 meeting in Turku

31.08.2022

Preparation of updated annual report for 2021

- Finalization of updated annual report for 2021

- Things to add
  - How many are software development papers
  - How many are method development papers
  - Source code availability

- Future plans with collected publication dataset

Manuscript draft about ML software for human microbiome analysis

- Finish sections before September (WG1)
- Include a companies section by Alina and Marcus
- IMDEA will work in style and homogenize till finish October
- Piotr describe figure and small paragraphs about software development during years for what section and citation
- FAIR software in the discussion
- Send mail to authors with no link to the authors Mihail
- Open manuscript for the other COST action member - beginning November
- Submit in December

Future plans

- Topics to discuss
  - Updated annual report (D1.5) for the year 2022.
  - Report outlining priority areas for new ML/statistics methods (D1.7) – will be replaced with opinion paper.