EMBnet.journal
Bioinformatics in Action

Ibrahimi et al. (2022) EMBnet.journal 27, e1012
http://dx.doi.org/10.14806/ej.27.0.1012

Reports

Report of the ML4Microbiome workshop 2021 - Statistical and Machine Learning Techniques for Microbiome Data Analysis

Eliana Ibrahimi[1,2], Ilze Elbere[3], Magali Bertland[4], Domenica D'Elia[5]✉

[1]Department of Biology, University of Tirana, Tirana, Albania
[2]Biomedical Research Institute, Hasselt University, Hasselt, Belgium
[3]Latvian Biomedical Research and Study Centre and University of Latvia, Riga, Latvia
[4]Université Paris-Saclay, INRAE, MGP, 78350 Jouy-en-Josas, France
[5]CNR, Institute for Biomedical Technologies, Bari, Italy
Competing interests: EI none; IE none; MB none; DD none

Abstract

The bacteria living in and on us, over 100 trillion, are essential for human development, immunity and nutrition and ultimately for human health. Researchers awareness of the importance of the microbiome for human wellbeing brought in the latest decade to the flourishing of new and numerous projects focused on the human microbiome, increasing the number of large microbiome datasets available. As a result, statistical and machine learning techniques to solve microbiome data analysis challenges are in high demand. The workshop "Statistical and Machine Learning Techniques for Microbiome Data Analysis" was organised by the COST Action ML4Microbiome to introduce the main concepts of study design and statistical/machine learning techniques used in human microbiome studies to a broad community of researchers and to foster connections between the discovery-oriented microbiome and statistical/machine learning researchers inside and outside the ML4Microbiome COST Action. To this aim, the workshop was organised as part of the training programme of the GOBLET & EMBnet Annual General Meeting 2021. This allowed ML4Microbiome to reach a broad and multidisciplinary community of scientists working in Bioinformatics and Computational Biology research and education from all over the world. The workshop attracted more than 80 participants from 40 different countries. In this paper, we report about the main topics treated and discuss the attendants' feedback regarding their level of satisfaction and their specific needs/demands for possible improvement of the training offer of ML4Microbiome in the microbiome research field. Under the authors' permission, presentations were recorded and are available as an ML4Microbiome playlist on YouTube (www.youtube.com/playlist?list=PL2aQNGWO1UUUmK7GJBK29yZr8Hnq6xFSD). The programme and presentations files are available at the ML4Microbiome website (https://www.ml4microbiome.eu/activities-and-events/ml4microbiome-workshop-on-statistical-and-machine-learning-techniques-for-microbiome-data-analysis/).

Workshop programme, aims and content

ML4Microbiome[1] - Statistical and machine learning techniques in human microbiome studies - is a COST Action (CA18131)[2] started in 2019 as an initiative of 24 European countries (currently 34) to establish a network aiming to improve the impact of microbiome research in life science research. To this aim, CA18131 specifically works to optimise and standardise statistical and machine learning techniques to analyse microbiome data (Moreno-Indias *et al.*, 2021; Marcos-Zambrano *et al.*, 2021). The correct usage of these techniques will allow researchers to better identify predictive and discriminatory 'omics' features, improve study reproducibility, and provide mechanistic insights into possible causal or contributing roles of the microbiome in human health and diseases.

[1]https://www.ml4microbiome.eu/ml4-microbiome-overview/
[2]https://www.cost.eu/actions/CA18131/

Article history
Received: 11 January 2022
Published: 2022

The workshop "Statistical and Machine Learning Techniques for Microbiome Data Analysis"[3] was held on October 15th and organised in the context of the GOBLET & EMBnet Annual General Meeting - Bioinformatics Education & Training from 2012 and beyond[4]. The organisation of the workshop in the context of this main event allowed ML4Microbiome to reach a vast audience of scientists in different disciplines and from many other countries, to foster connections between discovery-oriented microbiome researchers and statistical and machine learning experts inside and outside the ML4Microbiome COST Action. The GOBLET & EMBnet AGM 2021 was organised in collaboration with also the ELIXIR Italian Node[5].

Moreover, another event organised jointly with the leading conference was an ML4Microbiome Symposium entitled "Grand Challenges of Data-Intensive Science in microbiome & metagenome data analysis and training"[6]. This Symposium was held the day before the workshop (October 14th) and, the aims were, in addition to bringing to a large audience the latest advancements in the application of ML techniques to microbiome data in different research domains, also to serve as an introduction to the themes planned to be illustrated and discussed during the workshop.

The workshop's programme[3] was conceived to be of interest to a broad audience. The specific objective was to introduce attendants to the main challenges and possible solutions for optimising the experimental design and computational microbiome data analysis.

Domenica D'Elia, Leader of the ML4Microbiome Working Group 4[7], and co-organiser of the workshop along with Eliana Ibrahimi introduced the workshop programme, the trainers' role and expertise and provided a brief description of ML4Microbiome[8].

Ilze Elbere, a researcher in molecular biology at the Latvian Biomedical Research and Study Centre (Latvia) and lecturer at the University of Latvia (Latvia) who is also currently coordinating the Latvian Microbiome project (citizen science-based initiative), presented and discussed challenges and potential pitfalls of microbiome study design, sampling and wet-lab key steps. Various aspects of each topic were presented, including a short overview of the pros and cons of currently the most often used microbiome analysis methods, as well as examples from the lecturer's practical experience and available published data. The lecture also contained information on additional resources helpful when planning and conducting microbiome studies. The slides of the Ilze presentation are available on the ML4Microbiome website[9].

Eliana Ibrahimi, lecturer and researcher in biostatistics at the University of Tirana (Albania) and research affiliate at Hasselt University (Belgium), discussed the applications of statistics in microbiome data analysis using R statistics software. The first part of the lecture addressed the microbiome data structure and exploration. In this part, several exploration techniques applied to explore the microbiome were discussed with the R packages that implement them, such as Phyloseq (McMurdie and Holmes, 2013). This part was followed by the univariate (parametric and nonparametric) and multivariate community analysis applications in microbiome studies. Particular attention in this part was given to the multivariate analysis of variance with permutation (PERMANOVA) (Anderson, 2017) and the analysis of group similarities (ANOSIM). Then the compositional nature of microbiome data and ways to deal with it are briefly discussed. The last part of the lecture introduced several statistical models that can successfully be applied to model microbiome data. This part discussed the application of over-dispersed and zero-inflated models (Xia *et al.*, 2018), Dirichlet-multinomial models, zero-inflated longitudinal models (Fang *et al.*, 2016), and multivariate Bayesian mixed-effects models (Grantham *et al.*, 2020) in microbiome data modelling. The implementation in R and practical examples are introduced for each model above. The slides of this presentation are available here[10].

Magali Berland, leader of the ML4Microbiome Working Group 3 'Optimisation and Standardization' conducted a lecture on "How Artificial Intelligence is Enhancing Human Microbiome Research". Magali Berland is a research scientist in artificial intelligence and data science for the microbiome working at MetaGenoPolis[11], an INRAE unit composed of experts in gut microbiome research applied to health and nutrition. Her lecture introduced the three types of machine learning (unsupervised learning, supervised learning and reinforcement learning) and their main mechanisms. The current application of these techniques on microbiome data has been illustrated with many examples drawn from the recent literature. Finally, the current challenges and active research areas have been outlined. The slides of this presentation are available here[12].

[3]https://www.ml4microbiome.eu/activities-and-events/ml4microbiome-workshop-on-statistical-and-machine-learning-techniques-for-microbiome-data-analysis/

[4]https://embnet.eu/blog.html

[5]elixir-europe.org/about-us/who-we-are/nodes/italy

[6]https://www.ml4microbiome.eu/activities-and-events/ml4microbiome-symposium-grand-challenges-of-data-intensive-science-in-microbiome-metagenome-data-analysis-and-training-14-oct-2021/

[7]https://www.ml4microbiome.eu/working-group-descriptions/

[8]https://www.ml4microbiome.eu/wp-content/uploads/2021/10/ML4-welcome-at-ML4-WORKSHOP-15-oct-2021-DDElia.pdf

[9]https://www.ml4microbiome.eu/wp-content/uploads/2021/10/Microbiome_data_overview_ML4Microbiome_symposium_15.10.21_I.Elbere.pdf

[10]https://www.ml4microbiome.eu/wp-content/uploads/2021/11/Statistical-Analysis-of-Microbiome-Data-with-R-Eliana-Ibrahimi.pdf

[11]mgps.eu

[12]www.ml4microbiome.eu/wp-content/uploads/2021/10/AI-for-microbiome-research-Magali_Berland.pdf

Ibrahimi et al. (2022) EMBnet.journal 27, e1012
http://dx.doi.org/10.14806/ej.27.0.1012

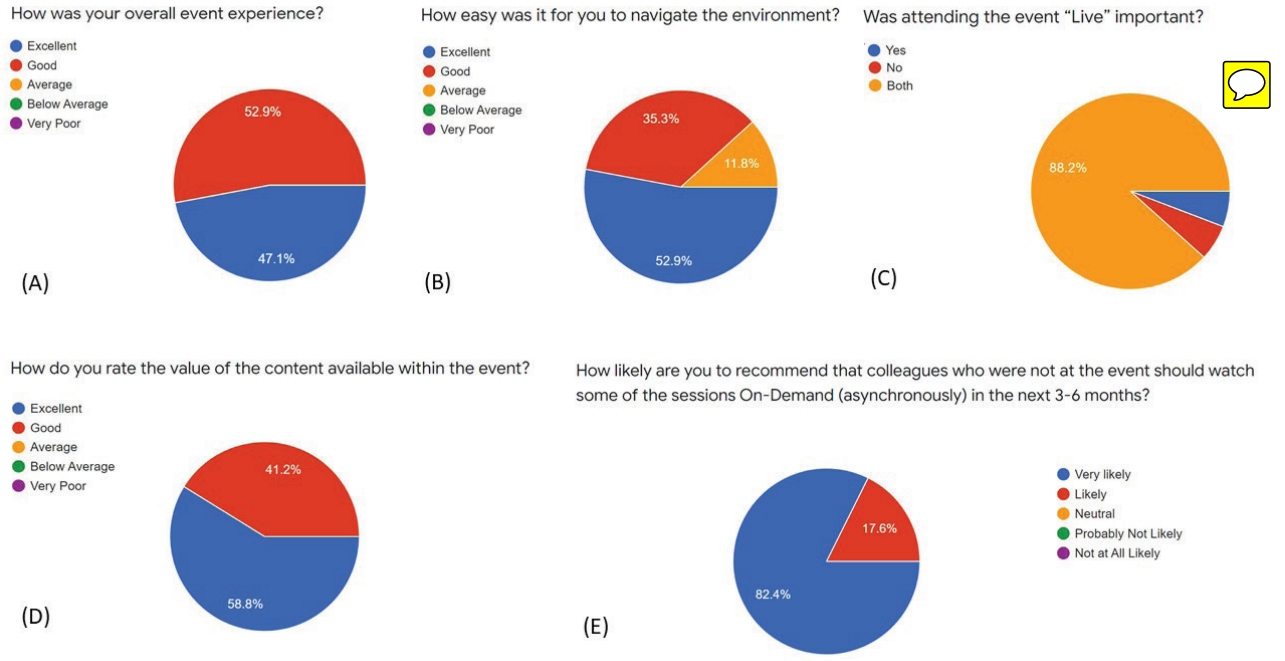

Figure 1. The figure shows the pie charts of post-workshop survey results. The survey was created using Google Drive Survey Forms facilities.

Workshop facilities, participation and evaluation

Due to the COVID-19 pandemic, the workshop was held as an online event using the Zoom conference web platform. A Google Drive shared folder was dedicated to the event to provide participants free access to the workshop material (programme and presentations) before and during the workshop. After the workshop, the speaker presentations were made public throughout the ML4Microbiome website[3] for to be visualised online or downloaded. Presentations from speakers were also recorded (under their permission) during the workshop, and videos are currently available as a YouTube playlist (ML4Microbiome Workshop 2021[13]) and as ML4Microbiome training material on the Action's website[14].

The registration to the event was free of charge but mandatory for organisational reasons, such as the need for moderation of access to the Zoom platform and apriori evaluation of the type of audience the workshop was going to have.

The number of people attending the workshop was 84 (including the speakers). ML4Microbiome members constituted 25%, others were from ELIXIR, EMBnet, GOBLET, and diverse Universities. Looking at attendants' geographical distribution and academic title, we have observed significant participation of young people (PhD or PostDoc) and broad geographical distribution. Many attendants were from Europe, but consistent involvement was also observed by Australia, Brazil, Ghana, India, Japan, Morocco, Mauritius, Pakistan, Perù, Sri Lanka and the USA.

After the workshop, attendants were invited to participate in a survey to collect their comments on the workshop's content and format and suggestions for improvements. People who participated in the survey were 20% of the total attendants. The main results are shown in Figure 1. The workshop experience was quoted as "Excellent" by 47,1% and as "Good" by 52,9% of attendants (Fig. 1-A). Rating the workshop content, 58,8% of participants judged it "Excellent", and 41,2% "Good" (Fig. 1-D); more than 80% claimed to be favourable to suggest colleagues not attending the workshop, to look at recorded sessions on-demand (Fig. 1-E). At the question: "From a content perspective, what could be improved?" (not shown in Fig. 1), the majority asked for more hands-on workshops.

Conclusions

From the point of view of the organisation of training initiatives, this event has provided valuable suggestions for the organisation of similar events in the upcoming year that can offer more hands-on sessions in experimental design and statistical analysis. On the contrary to what was believed in the pre-COVID era, researchers prefer online training events and appreciate the availability of recorded sessions (high demand) to be re-watched or first accessed after the event has taken place live. There were indeed many requests before and after the workshop for the availability of recorded

[13] https://www.youtube.com/playlist?list=PL2aQNGWO1UUUmK7GJBK29yZr8Hnq6xFSD
[14] https://www.ml4microbiome.eu/training-material-2/

presentations. If users' advantages are obvious, we learned with this experience that this training format also represents an advantage for trainers whose work is not just exhausted in the space of a few hours but can have a long-term value for the whole community of researchers.

Acknowledgements

This article is based upon work from COST Action CA18131 ML4Microbiome, supported by COST (European Cooperation in Science and Technology).

References

1. AAnderson MJ (2017) Permutational Multivariate Analysis of Variance (PERMANOVA). In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). http://dx.doi.org/10.1002/9781118445112.stat07841
2. Fang R, Wagner BD, Harris JK, Fillon SA (2016) Zero-inflated negative binomial mixed model: an application to two microbial organisms important in oesophagitis. Epidemiol Infect. **144**(11):2447-55. http://dx.doi.org/10.1017/S0950268816000662
3. Grantham NS, Guan Y, Reich BJ, Borer ET, Gross K (2020) MIMIX: A Bayesian Mixed-Effects Model for Microbiome Data From Designed Experiments. Journal of the American Statistical Association, **115**(530):599-609. http://dx.doi.org/10.1080/01621459.2019.1626242
4. Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V *et al.* (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. http://dx.doi.org/10.3389/fmicb.2021.634511
5. McMurdie PJ, Holmes S (2013) Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One **8**(4):e61217. http://dx.doi.org/10.1371/journal.pone.0061217
6. Moreno-Indias I, Lahti L, Nedyalkova M, Elbere I, Roshchupkin G *et al.* (2021) Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. Front Microbiol., **12**:635781. http://dx.doi.org/10.3389/fmicb.2021.635781
7. Xia Y, Sun J, Chen DG (2018) Statistical Analysis of Microbiome Data with R. ICSA Book Series in Statistics. Springer, Singapore.