Received: 27 September 2022 | Accepted: 28 September 2022

DOI: 10.1111/1462-2920.16232

CRYSTAL BALL

environmental microbiology Applied Microbiology International

Environmental microbiology going computational—Predictive ecology and unpredicted discoveries

Sara Hallin

Swedish University of Agricultural Sciences, Department of Forest Mycology and Plant Pathology, Uppsala, Sweden

Correspondence
Sara Hallin, Swedish University of Agricultural Sciences, Department of Forest Mycology and Plant Pathology, Uppsala, Sweden.
Email: sara.hallin@slu.se

Funding information
Sveriges Lantbruksuniversitet, Grant/Award Number: 2019-2024; Vetenskapsrådet, Grant/Award Number: 2016-03551

INCREASED DATA GENERATION AND DATA CRUNCHING

The fields of microbial ecology and environmental microbiology are producing loads of data, mainly nucleic acid sequence data due to the extensive use of amplicon sequencing and metagenomics, and an increasing use of transcriptomics. To increase our understanding of microorganisms in terrestrial ecosystems, multiple, concerted efforts to collect large numbers of samples for analyses of microbial communities were initiated already more than 15 years ago (Fierer & Jackson, 2006; Lozupone & Knight, 2007) but have really exploded the last years, with The Earth Microbiome Project Consortium being one of the first major endeavours for bacteria across all biomes (Thompson et al., 2017) and the work by Tedersoo et al. (2014) for soil fungi. The majority of the investigations have a biogeography focus based on a single sampling occasion and the word 'global' is frequently used in the titles of these soil microbial catalogues and surveys (Bahram et al., 2018; Delgado-Baquerizo et al., 2018; Gobbi et al., 2022). Similar efforts have been done for many other biomes. Although largely descriptive, they have contributed to a better understanding of microbial diversity and the distribution of microbial taxa and their functions at an unprecedented spatial scale. Further, correlative analyses have indicted direct or indirect drivers of the observed patterns as well as the role of microbial communities for ecosystem functioning (Bahram et al., 2018; Delgado-Baquerizo et al., 2020; Garland et al., 2021).

The massive amount of complex data is not only an opportunity but also a major challenge when it comes to meaningful interpretation. The field of computational biology, being the intersection of computer science and biology, is rapidly expanding and developing new methods for this purpose. Artificial intelligence (AI), including machine learning (ML) and to some extent also deep learning (DL) methods are promising for dealing with big data in microbial ecology and environmental microbiology (Ghannam & Techtmann, 2021; McElhinney et al., 2022). Especially ML approaches are increasingly adopted by ecologists and many of these methods will soon become routine tools for analyses of complex microbial omics data. They can be used to categorize and finds patterns in uncategorized data as well as analyse data that we know how to categorize. There are several advantages to using ML methods in microbiome studies, for example, they can deal with non-linear relationships, make better use of the full depth of high-dimensional data, and can be used to build predictive models based on environmental and community data.

Predictive modelling is very attractive in microbial ecology. Among the ML methods, random forests have become frequently applied in microbiome studies in the last decade (Jones et al., 2014; Ryo & Rillig, 2017). It is predominantly used for the identification of the best predictors for a given response variable and has for example been used to rank the environmental variables determining the major microbial phyla in wetlands (Bahram et al., 2022) and the diversity of ammonia oxidizing archaea across European soils (Saghaï

et al., 2022), as well as the relative importance of biotic and abiotic controls of nitrous oxide emissions from agricultural soils (Jones et al., 2022). Random forest modelling can be very useful when studying remote areas that are difficult to sample, as exemplified by climate projections on microbial communities in the Antarctic Ocean (Tonelli et al., 2021). RF models can also show how predictions change over the range of each individual predictor variable, thereby giving the possibility to identify thresholds or tipping points (Apley & Zhu, 2020; Saghaï et al., 2022). Already in 2012, artificial neural networks were used to incorporate interactions among community members in models for predictions of microbial community composition in time and space based on environmental data (Larsen et al., 2012). A similar approach was used to predict the maize rhizosphere community at different plant development stages or growth conditions (García-Jiménez et al., 2021). This type of approach can potentially assist in the microbiome engineering of important crops. However, with sequencing costs being relatively cheap, there is an increasing interest in using AI and microbiome data for microbiome-based diagnostics as a means to address environmental challenges and advance management practices (McElhinney et al., 2022). Two recent examples of the latter are the use of soil microbiome data to predict the propensity for specific plant diseases in agriculture (Yuan et al., 2020) and soil health metrics (Wilhelm et al., 2022), which can be laborious and expensive to measure. Combining ML and microbiome data has further shown promising in environmental monitoring, tracing of contaminants and predictions of environmental quality (Sperlea et al., 2022; Techtmann & Hazen, 2016; Wheeler, 2019), which allows us to move away from indicator taxa or microbial biomarkers and instead use the full breath of information encompassed by the microbial community in a given site or sample.

RE-USING DATA AND SHIFTING TO A DATA-DRIVEN COMPUTATIONAL SCIENCE

The large amounts of genetic data and corresponding meta-data generated in microbiome studies are real treasures, especially when it comes to metagenomes and metatranscriptomes, and only a fraction of the information available has been explored. This data can be used for meta-analyses to increase the scale of the study, but more importantly, it can be used to address other questions than those posed by the researchers that collected the original data. Making use of already published genome or sequence data in microbial ecology is not a new idea (Jones & Hallin, 2010) but now we have increasing possibilities to mine extremely large data sets (Coelho et al., 2022). Even more exciting are the possibilities to combine different types of data and information to go beyond the microbiome data. Integration of knowledge from diverse fields of research and the combination of microbiome data with other data from different sources have the potential to result in unexpected and unpredictable results, as well as new discoveries.

A recent example of re-using and combining data is the work by Ke et al. (2022), who reanalyzed data in published datasets on the effects of pesticide application on soil microbial communities combined with information on the physical and chemical properties of the pesticides. By developing a ML model, they were able to show that physical pesticide properties largely explain the ecological impact of the pesticide. This information can guide the design of pesticide molecules to minimize environmental risk. In the field of precision agriculture, researchers have proposed the integration of AI and nanotechnology with disparate datasets to enable the design of nanoscale agrochemicals for sustainable food production (Zhang et al., 2021). In another study, geographic and meteorological data as well plant-traits, land-use type and microbial community data were used in a ML-based prediction of grassland degradation, which is a multi-factorial phenomenon not easily captured by a few variables (Yan et al., 2022). Combining datasets and using computational approaches can also be used to develop new diagnostic tools. For example, de Andrade et al. (2021) suggest the development of a soil quality index based on soil microbiome data, crop productivity and a range of abiotic environmental factors to improve crop production systems using AI. Data-driven research relying on large, multiple, complex datasets and computational methods and capacity, as exemplified above, indicates a new paradigm in microbial ecology, and ecology in general (McCallen et al., 2019). We can anticipate new insights, similar to the leaps taken after advanced bioinformatics and multi-omics approaches became an integral part of microbial ecology research.

Microbial ecology and environmental microbiology will follow the trajectory in life sciences and become increasingly computationally demanding, focusing on larger and also more complex sets of information. We are already seeing the laboratories being sparsely populated while students, postdocs, and researchers spend increasing amount of time in front of their computers organizing and analysing data. My crystal ball says that a shift towards a data-driven rather than an experimental-driven and data generating science, that depends on complex, big data, and advanced technologies, will be a game changer in microbial ecology and environmental microbiology. This development is already putting pressure on management, storage and sharing of data. Data-driven microbial ecology research where different types of data are combined to consider the multidimensionality of ecosystems further suggests

that students and researchers not only need to enhance their computational skills, but also skills in working interdisciplinary. Nevertheless, important discoveries should ideally be followed by experimental approaches to test hypothesis, determine causal relationships, and verify mechanisms. Already, experimental validation is definitely a bottleneck to close the circle in microbial ecology research and, although my crystal ball is a bit hazy here, it looks like this will become an even greater bottleneck in the era of big data and data-drivenresearch in microbial ecology.

ACKNOWLEDGEMENTS

This work was supported by the Swedish University of Agricultural Sciences (senior career grant 2019-2024) and the Swedish Research Council (grant 2016-03551).

DATA AVAILABILITY STATEMENT

There is no data associated with this article.

ORCID

Sara Hallin https://orcid.org/0000-0002-9069-9024

REFERENCES

Apley, D.W. & Zhu, J. (2020) Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 82, 1059–1086.

Bahram, M., Espenberg, M., Pärn, J., Lehtovirta-Morley, L., Anslan, S., Kasak, K. et al. (2022) Structure and function of the soil microbiome underlying N_2O emissions from global wetlands. *Nature Communications*, 13, 1430.

Bahram, M., Hildebrand, F., Forslund, S.K., Anderson, J.L., Soudzilovskaia, N.A., Bodegom, P.M. et al. (2018) Structure and function of the global topsoil microbiome. *Nature*, 560, 233–237.

Coelho, L.P., Alves, R., del Río, Á.R., Myers, P.N., Cantalapiedra, C. P., Giner-Lamia, J.Q. et al. (2022) Towards the biogeography of prokaryotic genes. *Nature*, 601, 252–256.

Delgado-Baquerizo, M., Guerra, C.A., Cano-Díaz, C., Egidi, E., Wang, J.T., Eisenhauer, N. et al. (2020) The proportion of soil-borne pathogens increases with warming at the global scale. *Nature Climate Change*, 10, 550–554.

Delgado-Baquerizo, M., Oliverio, A.M., Brewer, T.E., Benavent-González, A., Eldridge, D.J., Bardgett, R.D. et al. (2018) A global atlas of the dominant bacteria found in soil. *Science*, 359, 320–325.

de Andrade, V.H.G.Z., Redmile-Gordon, M., Barbosa, B.H.G., Dini Andreote, F., Roesch, L. F.W. & Pylro, V.S. (2021) Artificially intelligent soil quality and health indices for 'next generation' food production systems. *Trends in Food Science & Technology*, 107, 195-200.

Fierer, N. & Jackson, R.B. (2006) The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 626–631.

García-Jiménez, B., Muños, J., Cabello, S., Medina, J. & Wilkinson, M.D. (2021) Predicting microbiomes through a deep latent space. *Bioinformatics*, 37, 1444–1451.

Garland, G., Edlinger, A., Banerjee, S., Degrune, F., García-Palacios, P., Pescador, D.S. et al. (2021) Crop cover is more important than rotational diversity for soil multifunctionality and cereal yields in European cropping systems. *Nature Food*, 2, 28–37.

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.

Gobbi, A., Acedo, A., Imam, N., Santini, R.G., Ortiz-Alvarez, R., Ellegaard-Jensen, L. et al. (2022) A global microbiome survey of vineyard soils highlights the microbial dimension of viticultural terroirs. *Communications Biology*, 5, 241.

Jones, C.M. & Hallin, S. (2010) Ecological and evolutionary factors underlying global and local assembly of denitrifier communities. *The ISME Journal*, 4, 633–641.

Jones, C.M., Putz, M., Emmerich, M. & Hallin, S. (2022) Reactive nitrogen restructures and weakens microbial controls of soil N_2O emissions. *Communications Biology*, 5, 273.

Jones, C.M., Spor, A., Brennan, F.P., Breuil, M.C., Bru, D., Lemanceau, P. et al. (2014) Recently identified microbial guild mediates soil N_2O sink capacity. *Nature Climate Change*, 4, 801–805.

Ke, M., Xu, N., Zhang, Z., Qui, D., Kang, J., Lu, R. et al. (2022) Development of a machine-learning model to identify the impacts of pesticides characteristics on soil microbial communities from high-throughput sequencing data. *Environmental Microbiology*. https://doi.org/10.1111/1462-2920.16175

Larsen, P.E., Field, D. & Gilbert, J.A. (2012) Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods*, 9, 621–625.

Lozupone, C.A. & Knight, R. (2007) Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 11436–11440.

McCallen, E., Knott, J., Nunez-Mir, G., Taylor, B., Jo, I. & Fei, S. (2019) Trends in ecology: shifts in ecology research themes over the past four decades. *Frontiers in Ecology and the Environment*, 17, 109–116.

McElhinney, J.M.W.R., Catacutan, M.K., Mawart, A., Hasan, A. & Dias, J. (2022) Interfacing machine learning and microbial omics: a promising means to address environmental challanges. *Frontiers in Microbiology*, 13, 852450.

Ryo, M. & Rillig, M.C. (2017) Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere*, 11, e01976.

Saghaï, A., Banerjee, S., Degrune, F., Edlinger, A., García-Palacios, P., Garland, G. et al. (2022) Diversity of archaea and niche preferences among putative ammonia-oxidizing Nitrososphaeria dominating across European arable soils. *Environmental Microbiology*, 24, 341–356.

Sperlea, T., Schenk, J.P., Dreßler, H., Beisser, D., Hattab, G., Boenigk, J. et al. (2022) The relationship between land cover and microbial community composition. *Science of the Total Environment*, 825, 153732.

Techtmann, S.M. & Hazen, T.C. (2016) Metagenomic applications in environmental monitoring and bioremediation. *Journal of Industrial Microbiology & Biotechnology*, 43, 1345–1354.

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N.S., Wijesundera, R. et al. (2014) Global diversity and geography of soil fungi. *Science*, 346, 1256688–1256682.

Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J. et al. (2017) A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, 551, 457–463.

Tonelli, M., Signori, C.N., Bendia, A., Neiva, J., Ferrero, B., Pellizari, V. et al. (2021) Climate projections for the Southern Ocean reveal impacts in the marine microbial communities following increases in sea surface temperature. *Frontiers in Marine Science*, 8, 636226.

Wheeler, N.E. (2019) Tracing outbreaks with machine learning. *Nature Reviews Microbiology*, 17, 269–269.

Wilhelm, R.C., van Es, H.M. & Buckley, D.H. (2022) Predicting measures of soil health using microbiome supervised machine learning. *Soil Biology and Biochemistry*, 164, 108472.

Yan, H., Ran, Q., Hy, R., Xue, K., Zhang, B., Zhou, S. et al. (2022) Machine learning-based prediction for grassland degradation

using geographic, meteorological, plant and microbial data. *Ecological Indicators*, 137, 108738.

Yuan, J., Wen, T., Zhang, H., Zhao, M., Penton, C.R., Thomashow, L. S. et al. (2020) Predicting disease occurrence with high accuracy based on soil macroecological patterns of Fusarium wilt. *The ISME Journal*, 14, 2936–2950.

Zhang, P., Guo, Z., Ullah, S., Melagraki, G., Afantis, A. & Lynch, I. (2021) Nanotechnology and artificial intelligence to enable sustainable and precision agriculture. *Nature Plants*, 7, 864–876.

How to cite this article: Hallin, S. (2023) Environmental microbiology going computational—Predictive ecology and unpredicted discoveries. *Environmental Microbiology*, 25(1), 111–114. Available from: https://doi.org/10.1111/1462-2920.16232