EcoPLOT: dynamic analysis of biogeochemical data

Permalink
https://escholarship.org/uc/item/9h33q1m0

Journal
Bioinformatics, 38(5)

ISSN
1367-4803

Authors
Sanchez, Christopher D
Brown, J Benjamin
Gal-Oz, Omree
et al.

Publication Date
2022-02-07

DOI
10.1093/bioinformatics/btab842

Peer reviewed
1 Introduction

Microorganisms are involved in global and local biogeochemical cycles and redox reactions that both directly and indirectly affect the functioning of their surroundings (Maier, 2015; Wiedler et al., 2015). In soils, microbes are direct promoters of plant health and development through their participation in a number of redox reactions, promotion of nutrient cycling and conference of resistance to biotic and abiotic stresses (Miransari, 2014). Bacterial, archaeal and fungal community compositions are often heterogeneous across soils and are impacted by fluctuations in environmental conditions such as weather and soil chemistry (Serna-Chavez et al., 2013). The environmental variability and complexity challenge our understanding of processes such as soil ecosystem functioning.

To study ecosystem processes, high-throughput molecular techniques, such as DNA/RNA sequencing, metabolomics and soil chemistry analyses, which require destructive sampling, are regularly combined with in situ measuring devices, for example using real-time minimally or non-invasive sensors and cameras (Singer et al., 2021). Depending on measurement frequency and choice of instrumentation, these approaches can result in large amounts of data of different formats. The resulting datasets are often large and complex and demand proper data integration as well as iterative, custom and dynamic analysis tools.

Existing tools for microbial analysis, such as QIIME (Caporaso et al., 2010), QIIME 2 (Bolyen et al., 2019) and Phylloseq (McMurdie and Holmes, 2013), provide reproducible workflows for the analysis of amplicon data and feature the ability to demultiplex sequences, assign taxonomies and perform various downstream statistical analyses. However, customization of functionality is usually limited to the tool developers, while the existing code may not satisfy the unique needs of a study design, hence requiring additional custom scripts. Furthermore, these tools were designed for microbiome-centric studies that do not allow separate analysis of environmental datasets, such as soil chemistry profiles and plant phenotyping data.

Here, we introduce EcoPLOT (parameterized linkage of omics-driven technologies), a web-based tool for the visualization and analysis of multivariate datasets, as well as its accompanying R package (R Core Team, 2021). EcoPLOT features an interactive graphical user interface which provides an analytical workflow that educates users on how to integrate and centrally analyze complex multi-disciplinary datasets, quickly discover significant interactions and trends within multifactorial datasets and to produce publication-ready visualizations. Besides providing an assembly of state-of-the-art analysis tools for biogeochemical datasets based on few established R libraries, EcoPLOT also offers a suggested workflow that educates and introduces users to analysis methods, such as novel application of the iterative random forest (iRF) to microbiome research. Improving our understanding of the reciprocal interactions between microbes, and the environments and hosts they reside in is essential for various environmental missions including improvements in crop yield, carbon sequestration and bioremediation.

2 Materials and methods

EcoPLOT provides users with state-of-the-art tools for graphical and statistical analysis of datasets that explore plant phenotypes, geochemistry and microbiome community dynamics, separately and in combination with one another. EcoPLOT’s interface is organized into four modules (Plant, Microbiome, Environment and iRF), each
Trends in plant and environmental modules can help to inform the identification of significant drivers using graphical and regression analyses as well as machine learning (iRF)

Besides standard regression methods, EcoPLOT offers users the iterative Random Forest (iRF) algorithm (Basu et al., 2018) to uncover significant groupings between plant, environment, and the microbiome. iRF searches for stable, high order interactions within biological datasets at minimal computational cost, which allows computation on a local computer. The iRF algorithm includes user-uploaded data into one large dataset that can be accessed directly through the Machine Learning ‘iRF’ module. After optional data subsetting, response variables can be selected to be tested. Testing and training datasets are created automatically following this selection. Output consists of Variable Importance Plots that display the top factors in descending order as measured by a Random Forest and a list of interacting variables and their corresponding stability score. Users are able to create surface plots to visualize how two interacting variables affect the desired response. EcoPLOT uses default iRF parameters (n.iter = 5, ntree = 500, n.bootstrap = 30), however, they can be changed directly within the R code to fit a given experiment.

EcoPLOT provides a comprehensive toolbox for the multivariate analysis of biogeochemical datasets and is open source, built almost entirely from R code. Like other shiny apps it can be further customized using HTML, CSS, JavaScript or additional R packages and can be expanded to support other types of metadata. Because of its streamlined functionality, data exploration with IRF is simple to set up and can be implemented for data of any size. Based on the parameters defined for iRF, run times may vary, however EcoPLOT can be expanded to support processing over an external server. All EcoPLOT functionality is described in detail in the Instructions and data type specific guides.

**Funding**

This material by the TWIN Ecostem project led by Lawrence Berkeley National Laboratory is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological Environmental Research under contract number DE-AC02-05CH11231 and under the LBNL LDRD entitled ‘Improving biofuel crop yield from field to lab under drought stress’ (to E.S.).
Conflict of Interest: none declared.

References


