Bioinformatics, 38(5), 2022, 1480–1482
https://doi.org/10.1093/bioinformatics/btab842
Advance Access Publication Date: 20 December 2021
Applications Note

Data and text mining

EcoPLOT: dynamic analysis of biogeochemical data

Christopher D. Sanchez [1,*], J. Benjamin Brown[1], Omree Gal-Oz[1] and Esther Singer[1,2,*]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94710, USA and [2]DOE Joint Genome Institute, Berkeley, CA 94720, USA

*To whom correspondence should be addressed.

Associate Editor: Zhiyong Lu

Received on June 2, 2021; revised on December 2, 2021; editorial decision on December 9, 2021; accepted on December 14, 2021

Abstract

Motivation: We have created EcoPLOT (parameterized linkage of omics-driven technologies), a web-app for the dynamic, interactive analysis of biogeochemical datasets that combines state-of-the-art analysis tools to statistically and graphically explore environmental, geochemical and microbiome datasets. Using the iterative random forest, a machine learning algorithm, EcoPLOT allows for the *de novo* discovery of drivers which exhibit significant impact on plant, microbial or soil dynamics.

Availability and implementation: EcoPLOT is built entirely within the R language. It can be accessed through any system where R is installed, including Windows, Mac and most Linux systems. EcoPLOT is free to use and can be accessed at https://github.com/cdsanchez18/EcoPLOT.

Contact: cdsanchez@lbl.gov or esinger@lbl.gov

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; Wieder *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 Phyloseq (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

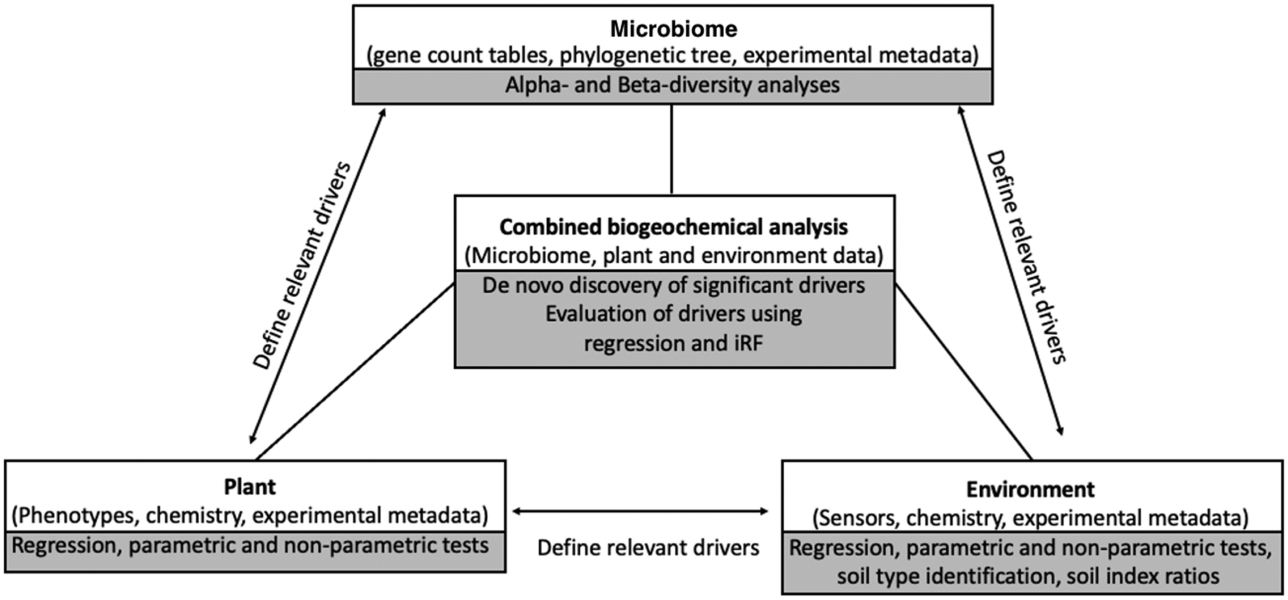
Microbiome
(gene count tables, phylogenetic tree, experimental metadata)
Alpha- and Beta-diversity analyses
Define relevant drivers
Define relevant drivers
Combined biogeochemical analysis
(Microbiome, plant and environment data)
De novo discovery of significant drivers
Evaluation of drivers using
regression and iRF
Plant
(Phenotypes, chemistry, experimental metadata)
Regression, parametric and non-parametric tests
Define relevant drivers
Environment
(Sensors, chemistry, experimental metadata)
Regression, parametric and non-parametric tests,
soil type identification, soil index ratios

Fig. 1. EcoPLOT workflow overview. Plant, environment and microbiome modules offer data-specific statistical and graphical analysis tools. Data tables can be merged into any module and be used for combined analysis for discovery of significant drivers using graphical and regression analysis as well as machine learning (iRF)

of which provide graphics and analysis options specific to the corresponding data type. We explain the unique features of each module, respective data types and a general workflow (Fig. 1) below.

The plant module 'Phenotype Data' allows import of spatially and temporally resolved plant phenotypic measurements. Within this module users will be able to identify relationships between plant phenotype and experimental design. The imported dataset is automatically summarized and quartiles for each relevant measurement are presented. The user can subset their dataset based on the inherent data ranges or sample factors. Statistical analyses include one-way and two-way ANOVA, Tukey HSD, two sample t-test and Kruskal–Wallis statistical methods. Data visualization includes boxplots, histograms and interactive scatter plots.

The environmental module 'Environmental Data' allows users to investigate spatial and temporal trends of any environmental parameters, including time-stamped sensor data and physical/chemical analyses. For categorical temporal and spatial variables, the same analyses as provided in the plant module can be applied to the environmental module. In addition, users with soil chemistry data can identify their soil type using an interactive soil texture triangle. Trends in plant and environmental modules can help to inform the microbiome data.

The microbiome module 'Microbiome Data' provides users with an extensive toolkit to analyze quality-screened sequence count tables e.g. amplicon sequence variant (ASV) tables, and associated metadata and phylogenetic trees. Data input formats are compatible with QIIME 1&2, Phyloseq, as well as non-formatted files (i.e. text files). Standard phylogenetic tree data formats are also accepted. Users can filter count tables by frequency across samples and # of counts per gene. Alpha (Shannon's H index, Chao1, observed OTUs/ASVs) and beta (Bray–Curtis, Jaccard, Euclidean, unweighted/weighted unifrac) diversity can be displayed using an array of visualization methods. Statistical analyses, such as differential abundance calculations using DESeq2 (Anders and Huber, 2010), Kruskal–Wallis and ADONIS (Analysis of variance using distance matrices) are applied on demand. Datasets are automatically reformatted to meet the requirements of their respective analyses.

Across all modules graphical analysis of scatter and principal coordinate analysis (PCO) plots is interactive, allowing users to identify outliers and create up to ten different custom group labels which can be accessed in downstream analyses, such as statistical tests or 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. EcoPLOT seamlessly integrates 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 easily 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

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol*, 11, R106.

Basu,S. *et al.* (2018) Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci. USA*, **115**, 1943–1948.

Bolyen,E. *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, **37**, 852–857.

Caporaso,J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Maier,R.M. (2015) Biogeochemical cycling. *Environmental Microbiology: Third Edition, 339–373*.

McMurdie,P.J. and Holmes,S. (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, **8**, e61217.

Miransari,M. (2014) Plant growth promoting Rhizobacteria. *J. Plant Nutr.*, **37**, 2227–2235.

R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Serna-Chavez,H.M. *et al.* (2013) Global drivers and patterns of microbial abundance in soil. *Glob. Ecol. Biogeogr.*, **22**, 1162–1172.

Singer,E. *et al.* (2021) Novel and emerging capabilities that can provide a holistic understanding of the plant root microbiome. *Phytobiomes J.* 5(2), 122–132.

Wieder,W.R. *et al.* (2015) Explicitly representing soil microbial processes in Earth system models: soil microbes in earth system models. *Glob. Biogeochem. Cycles*, **29**, 1782–1800.