The Journal of Open Source Software

PhenoFeatureFinder: a python package for linking developmental phenotypes to omics features

Lissy-Anne M. Denkers[1], **Marc D. Galland**[2], **Annabel Dekker**[3], **Valerio Bianchi**[3,4], **and Petra M. Bleeker**[1]

1 University of Amsterdam, Department of Plant Physiology, Green Life Science Research Theme, Swammerdam Institute for Life Sciences, Amsterdam, The Netherlands **2** INRAE, Institute of Genetics, Environment and Plant Protection (IGEPP—Joint Research Unit 1349), Le Rheu, France **3** Enza Zaden R&D B.V., BTR-BM Bioinformatics, Enkhuizen, The Netherlands **4** Wageningen Bioveterinary Research, Wageningen University & Research, Lelystad, Netherlands

DOI: 10.21105/joss.07264

Software

- Review
- Repository
- Archive

Editor: Julia Romanowska

Reviewers:

- @admiralenola
- @lopeztarifa
- @pierrepeterlongo

Submitted: 05 June 2024
Published: 20 November 2024

Summary

PhenoFeatureFinder is designed to facilitate the analyses required to analyse quantitative and/or progressive phenotypic- and omics data, and link those using Machine Learning with the aim to identify causal features, in one package. It can be used for 1) evaluation and visualisation of phenotype progression over multiple stages and between groups (e.g. treatments, genotypes), 2) pre-processing of omics data, and 3) prediction of features that explain the phenotypic classification. To facilitate usability, each step in the pipeline can also be performed independently, hence has been assigned a class in the package (Figure 1). We provide an example of implementation below that focuses on insect development through time and the selection of metabolic features causal to the observed phenotype, but different input data could be used, provided it has a similar structure. This could be any phenotype that is scored in progressive stages over time. Also, PhenoFeatureFinder was developed initially with metabolomics data, but users can evaluate its fit applying other types of omics data.

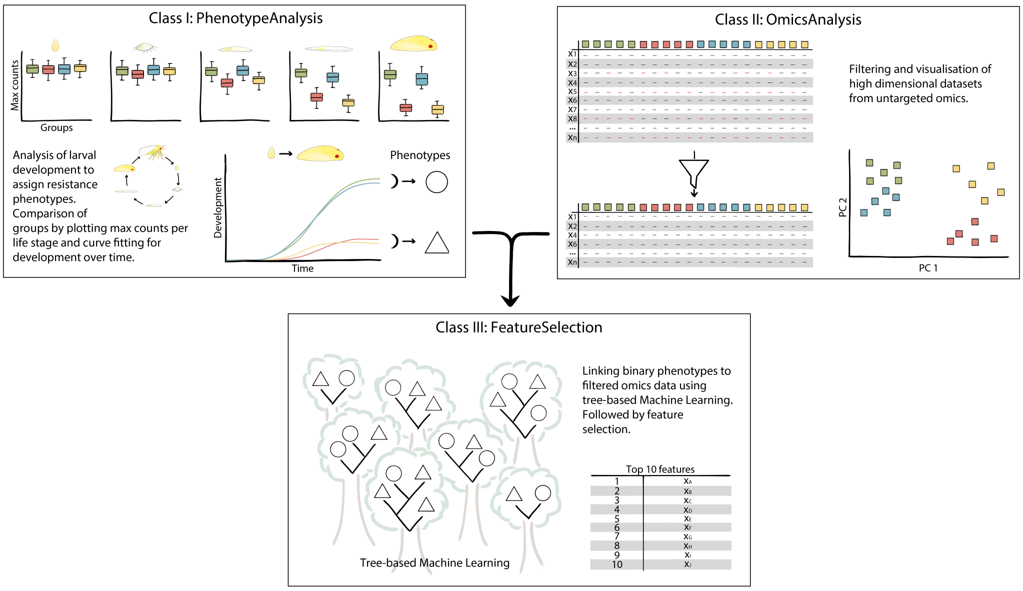
Class I: PhenotypeAnalysis
Max counts
Groups
Analysis of larval development to assign resistance phenotypes. Comparison of groups by plotting max counts per life stage and curve fitting for development over time.
Development
Time
Phenotypes
Class II: OmicsAnalysis
X1
X2
X3
X4
X5
X6
X7
X8
...
Xn
Filtering and visualisation of high dimensional datasets from untargeted omics.
X1
X2
X4
X6
...
Xn
PC 2
PC 1
Class III: FeatureSelection
Linking binary phenotypes to filtered omics data using tree-based Machine Learning. Followed by feature selection.
Tree-based Machine Learning
Top 10 features
1 XA
2 XB
3 XC
4 XD
5 XE
6 XF
7 XG
8 XH
9 XI
10 XJ

Figure 1: Overview of the package, consisting of three classes that can be used separately or as a workflow. Class 1: analysing and visualising the phenotype, Class 2: preprocessing and visualising omics datasets, and Class 3: feature selection through a Machine Learning approach.

Statement of need

The analysis of developmental phenotypes can be challenging, due to the many variables involved (e.g. time, developmental stages, replicates, treatments), especially for researchers whose strength or interest does not lie in data analysis. The same goes for the pre-processing of omics data and linking the omics data and developmental phenotypes. With `PhenoFeatureFinder`, we aim to support such research by combining the nessecary functionalities in one package with easy to follow manuals and examples.

In R, the package `drc` is available for fitting dose-response curves (Ritz et al., 2015), offering an extensive and versatile set of functionalities. However, for the purposes described here `drc` poses some limitations, such as the options for custom pre-processing and analyses of multiple experimental groups simultaneously. Here we implemented pre-processing steps and aimed to decrease the amount of coding needed to obtain a fitted development curve.

Use case example

Plants interact with their (a)biotic environment through a range of specialised metabolites and deal with pathogens and pest attack through constitutive or inducible production of those defence molecules (Erb & Kliebenstein, 2020; García-Olmedo et al., 1998). High-throughput "-omics" tools including (untargeted) metabolomics have been successfully implemented in plant biology (Dalio et al., 2021), but the accompanying resistance phenotyping often lacks in robustness (Song et al., 2021).

Proliferation of an insect population is affected by various factors, including the chemical composition of the host, and/or the environment (Ma et al., 2022). In particular, host resistance via hampered larval development is noteworthy, because reducing the speed at which larvae reach the adult stage and produce offspring negatively affects pest-population development (Maharijaya et al., 2019; Muema et al., 2016; Vengateswari et al., 2022). However, evaluating larval development results in a complex dataset that is challenging to process. Developmental success is based on the number of larvae throughout various larval stages, as well as on the speed of development.

To identify underlying mechanisms of resistance, the chemical or molecular composition of a plant can be investigated. Proteins and metabolites are commonly analysed through untargeted Mass-Spectrometry, yielding exhaustive profiles generally consisting of many thousands of unannotated features. Often such data displays sparsity, i.e. missing values between datasets, and a low sample-to-feature ratio, adding to the complexity of the analysis (Kortbeek et al., 2021; Liebal et al., 2020). Tree-based Machine-Learning algorithms (e.g., random forest) are suitable for the analysis of, and feature selection from, untargeted data (Liebal et al., 2020) computing the contribution of each feature in the phenotypic classification.

Class I: `PhenotypeAnalysis`

A binary classification of plants into "resistant" or "susceptible" helps to extract relevant features especially when threshold effects or sparsity (presence/absence) effects are at play. Here we firstly assess performance over different developmental stages of larvae on different host plants. The number of individuals in each stage at a given time is recorded. When plotted, the cumulative data of these bioassays resemble a growth- or dose-response curve that can be used to manually assign a binary phenotype (e.g., resistant/non-resistant), a resistance classification used as input for `FeatureSelection` (Class 3).

To account for missing data when individuals that reached the final developmental stage are removed from the experiment, we implemented an automated correction step. The count data can be transformed to cumulative data to analyse the maximum of individuals that reach each of the developmental stages. Next, the time to reach a specific stage can be compared between treatments by fitting a 3-parameter log-logistic curve (Muse et al., 2021; Seefeldt et al., 1995; Vliet & Ritz, 2013) to the cumulative data for each treatment, with the function:

$$f(x) = \frac{m}{1 + \exp(s \times (\log(x) - \log(e_{50})))}$$

where x is time, m is the upper limit (or maximum of individuals that developed to the stage of interest), s is the slope of the linear part of the curve and e_{50} is the EmT50 (the timepoint at which 50% of the individuals have developed to the stage of interest). We added the possibility to compare performance between treatments by fitting a curve with the function:

$$f(x) = \frac{a \times \frac{s}{m} \times (\frac{x}{m})^{s-1}}{1 + (\frac{x}{m})^{s}}$$

Here, x is time, a the area under the curve, s is the shape of the curve and m the median time point. Both functions output a table with the model parameters, confidence intervals and the model fit, together with a plot displaying the observed data and the fitted model. For both functions it is possible to predict the potential maximum beyond the final experimental measurements.

Class II: `OmicsAnalysis`

Untargeted omics results in large datasets that tend to contain background noise and unreliable features. To clean the data, multiple filtering methods are implemented in the `OmicsAnalysis` class, including the removal of contaminants present in blank samples, filtering to decrease sparsity and other quality control steps. The structure of the data can subsequently be visualised with a PCA and an UpSet plot.

Class III: `FeatureSelection`

Combining the output of Classes 1 and 2, i.e. the binary phenotype classification and the tidied untargeted metabolomics, `FeatureSelection` is set up to predict features that can explain the phenotypic observation under study. This part of the pipeline was built as a wrapper around the Python libraries `scikit-learn` and `TPOT` (Olson et al., 2016; Pedregosa et al., 2011). The `FeatureSelection` wrapper is designed to select optimal pipelines for data preprocessing and identification of the most suitable Machine Learning model. One characteristic of metabolomics data is strongly correlated features (linear dependencies between variables) that make it difficult to extract individual feature importance. Therefore, this method implements a PCA as dimensionality reduction method before searching for the best fitting pipeline. Finally, the importance of the Principal Components and their most related features (high loadings) can be retrieved to select features with predicted importance to the phenotypic classification.

Acknowledgements

The Amsterdam Data Science Centre is acknowledged for their input and Frans van der Kloet for his statistical support. Gea van der Lee and Piet Verdonschot are kindly thanked for providing the data from (Lee et al., 2020) used in the example in GitHub.

Author contributions

The software was written by Lissy-Anne Denkers (LD) and Marc Galland (MG), with input imput from Petra Bleeker (PB) and tested by Annabel Dekker (AD) and Valerio Bianchi (VB). The manuals and examples were written by LD with major input from AD and VB. The manuscript was designed, written and revised by LD, MG, AD, VB and PB.

References

Dalio, R. J. D., Litholdo, C. G., Arena, G., Magalhães, D., & Machado, M. A. (2021). *Contribution of omics and systems biology to plant biotechnology* (pp. 171–188). https://doi.org/10.1007/978-3-030-80352-0_10

Erb, M., & Kliebenstein, D. J. (2020). Plant secondary metabolites as defenses, regulators, and primary metabolites: The blurred functional trichotomy. *Plant Physiology*, *184*, 39–52. https://doi.org/10.1104/pp.20.00433

García-Olmedo, F., Molina, A., Alamillo, J. M., & Rodríguez-Palenzuéla, P. (1998). Plant defense peptides. *Biopolymers*, *47*, 479–491. https://doi.org/10.1002/(SICI)1097-0282(1998)47:6%3C479::AID-BIP6%3E3.0.CO;2-K

Kortbeek, R. W. J., Galland, M. D., Muras, A., Kloet, F. M. van der, André, B., Heilijgers, M., Hijum, S. A. F. T. van, Haring, M. A., Schuurink, R. C., & Bleeker, P. M. (2021). Natural variation in wild tomato trichomes; selecting metabolites that contribute to insect resistance using a random forest approach. *BMC Plant Biology*, *21*, 315. https://doi.org/10.1186/s12870-021-03070-x

Lee, G. H. van der, Kraak, M. H. S., Verdonschot, R. C. M., & Verdonschot, P. F. M. (2020). Persist or perish: Critical life stages determine the sensitivity of invertebrates to disturbances. *Aquatic Sciences*, *82*. https://doi.org/10.1007/s00027-020-0698-0

Liebal, U. W., Phan, A. N. T., Sudhakar, M., Raman, K., & Blank, L. M. (2020). Machine learning applications for mass spectrometry-based metabolomics. *Metabolites*, *10*, 243. https://doi.org/10.3390/metabo10060243

Ma, K., Tang, Q., Liang, P., Li, J., & Gao, X. (2022). A sublethal concentration of afidopyropen suppresses the population growth of the cotton aphid, aphis gossypii glover (hemiptera: aphididae). *Journal of Integrative Agriculture*, *21*, 2055–2064. https://doi.org/10.1016/S2095-3119(21)63714-0

Maharijaya, A., Vosman, B., Pelgrom, K., Wahyuni, Y., Vos, R. C. H. de, & Voorrips, R. E. (2019). Genetic variation in phytochemicals in leaves of pepper (capsicum) in relation to thrips resistance. *Arthropod-Plant Interactions*, *13*, 1–9. https://doi.org/10.1007/s11829-018-9628-7

Muema, J. M., Bargul, J. L., Nyanjom, S. G., Mutunga, J. M., & Njeru, S. N. (2016). Potential of camellia sinensis proanthocyanidins-rich fraction for controlling malaria mosquito populations through disruption of larval development. *Parasites & Vectors*, *9*, 512. https://doi.org/10.1186/s13071-016-1789-6

Muse, A. H., Mwalili, S. M., & Ngesa, O. (2021). On the log-logistic distribution and its generalizations: A survey. *International Journal of Statistics and Probability*, *10*, 93. https://doi.org/10.5539/ijsp.v10n3p93

Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016). Evaluation of a tree-based pipeline optimization tool for automating data science. *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, 485–492. https://doi.org/10.1145/2908812.2908918

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://dl.acm.org/doi/10.5555/1953048.2078195

Ritz, C., Baty, F., Streibig, J. C., & Gerhard, D. (2015). Dose-response analysis using R. *PLoS ONE*, *10*, 1–13. https://doi.org/10.1371/journal.pone.0146021

Seefeldt, S. S., Jensen, J. E., & Fuerst, E. P. (1995). Log-logistic analysis of herbicide dose-response relationships. *Weed Technology*, *9*, 218–227. https://doi.org/10.1017/s0890037x00023253

Song, P., Wang, J., Guo, X., Yang, W., & Zhao, C. (2021). High-throughput phenotyping: Breaking through the bottleneck in future crop breeding. *The Crop Journal*, *9*, 633–645. https://doi.org/10.1016/j.cj.2021.03.015

Vengateswari, G., Arunthirumeni, M., Shivaswamy, M. S., & Shivakumar, M. S. (2022). Effect of host plants nutrients, antioxidants, and phytochemicals on growth, development, and fecundity of spodoptera litura (fabricius) (lepidoptera: noctuidae). *International Journal of Tropical Insect Science*, *42*, 3161–3173. https://doi.org/10.1007/s42690-022-00868-6

Vliet, L. van der, & Ritz, C. (2013). Statistics for analyzing ecotoxicity test data. In J.-F. Férard & C. Blaise (Eds.), *Encyclopedia of Aquatic Ecotoxicology* (pp. 1081–1095). Springer Dordrecht. https://doi.org/10.1007/978-94-007-5704-2