Development of an R Toolbox for Near-Infrared Spectroscopy Data Processing and Analysis of Plant Metabolic Phenotypes

Deggendorf Institute of Technology

MSc. Life Science Informatics

Methun George

Supervised by
PD Dr. habil. rer. nat. Steffen Neumann
Prof. Dr. Melanie Kappelmann-Fenzl

December 25, 2024

Contents

Chapter 1

Introduction

The understanding of interplay between plant physiology and its hidden biochemical process is crucial for the improvement of basic plant science and addressing global challenges such as food security, crop resilience and combating climate change [1]. In recent years, advanced High-throughput analytical techniques such as Near-Infrared Spectroscopy (NIRS) and Liquid Chromatography-Mass Spectrometry (LC-MS) has instigated a paradigm shift in plant biology [2][3]. These High-throughput techniques are mostly used in areas like genomics, imaging and spectroscopy and is known for their ability to collect and analyse the data faster than traditional techniques[3]. High-throughput techniques are widely used since they enable the efficient collection of vast amount of data at various scales, from molecular to field level over significant time periods[4]. The big data generated by these high throughput procedures present both opportunities and challenges at the same time. It requires efficient processing to extract maximum useful results and this is where Machine Learning (ML) or Deep Learning (DL) becomes indispensable [4][5]. ML as a part of Artificial Intelligence (AI) refers to the ability of computers to find patterns and learn from the existing data which can be employed in processing high dimensional data [6][4]. The ML algorithms are powerful enough to analyse complex, high dimensional datasets, enabling accurate predictions of plant traits or other features based on the input data. Additionally, integrating these big data with ML could help the researchers to optimize data processing pipelines, enhance predictive accuracy and thereby enter into a new era of data-driven decision-making [4][5]. This project employs linear models, non-linear models and neural networks to predict various plant features and compare their performances.

A significant shift in the realm of the biomedical community has brought new guidelines to ensure readability, modularity, transparency and extensibility of computational toolboxes. A toolbox, which stores multiple functions, parameters and results in a central location should be maintainable and uncomplicated for the developers and members of the open-source community [7]. R is a powerful and widely used programming language in the analysis and processing of high throughput data. Additionally, R contains a multitude of statistical and high quality visualization packages such as ggplot2 which are capable of processing and integrating big data to different ML methods [8]. Bioconductor is an open source R software for bioinformatics, which contains more than 3000 packages for statistical computing. This offers an object oriented framework for the high dimensional data, cutting edge visualization capabilities and interoperability [9]. Existing tools in Near-Infrared Spectroscopy (NIRS) data processing lack functionalities that could simplify and standardize data workflows when integrated with the SummarizedExperiment framework from the Bioconductor package. To address these gaps, the R toolbox, "nearspectRa" was developed for processing NIRS data. This package has a modular structure which creates a SummarizedExperiment object from NIRS data.

Metabolomics, the study of small molecular compounds in biological systems, is a rapidly advancing field of science with applications in biotechnology, medicine, synthetic biology and environmental science [6]. Metabolomics has emerged as a transformative tool in plant biology, enabling cost-efficient and high throughput molecular characterization. The integration of metabolomics with different omics approaches has proven invaluable for functional genes identification and developing trait specific markers [10]. Metabolomics, which is built on the advancement of phenomics and genomics, provides high throughput and precise profiling of metabolites, revealing the physiological state of cells [6][10]. Metabolites play a crucial role in plant metabolism, influencing its biomass and architecture therefore study of these small molecules will aid in uncovering plant regulatory mechanisms and pathway interactions [10].

Chapter 2

Background

The background of this study include

2.1 Related Work

2.2 Near Infrared Spectroscopy (NIRS)

2.3 R Programming

2.4 Machine Learning

2.4.1 Partial Least Square Regression (PLSR)

2.4.2 Random Forest (RF)

2.4.3 Convolutional Neural Network (CNN)

2.5 Mass Spectrometry and Liquid Chromatography

Chapter 3

Implementation

3.1 Packages

Github, testing, actions

3.2 Contributions elsewhere

3.3 HPC runs

Chapter 4

Results and Discussion

4.1 Data charecterestics

histogram, spectra

4.2 Baseline Machine Learning Models Pablo

PLS, RF, CNN

4.2.1 Variable importance

4.3 Variations in Baseline systems

4.3.1 modifying the Test and Training split

4.3.2 input data length

4.4 Sues

Chapter 5

Reference

1. Pieruschka R, Schurr U. Plant Phenotyping: Past, Present, and Future. Plant Phenomics. 2019 Mar 26;2019:7507131. doi: 10.34133/2019/7507131. PMID: 33313536; PMCID: PMC7718630.
2. Pulok K. Mukherjee, Quality control and evaluation of herbal drugs, Evaluating natural products and traditional medicine. 2019, doi:10.1016/C2016-0-042328, ISBN:978-0-12-813374-3
3. Nizamani, M. M., Zhang, Q., Muhae-Ud-Din, G., Wang, Y. (2023). High-throughput sequencing in plant disease management: A comprehensive review of benefits, challenges, and future perspectives. Phytopathology Research, 5(44). https://doi.org/10.1186/s42483-023-00215-7
4. Lane, H. M., Murray, S. C. (2021). High throughput can produce better decisions than high accuracy when phenotyping plant populations. Crop Science, 61(3), 1473–1484. https://doi.org/10.1002/csc2.20514
5. Zhang, N., Zhou, X., Kang, M., Hu, B.-G., Heuvelink, E., Marcelis, L. F. M. (2023). Machine learning versus crop growth models: An ally, not a rival. Journal of Experimental Botany, 74(4), 1259–1276. https://doi.org/10.1093/jxb/erac517
6. Zhu H. Big Data and Artificial Intelligence Modeling for Drug Discovery. Annu Rev Pharmacol Toxicol. 2020 Jan 6;60:573-589. doi: 10.1146/annurev-pharmtox-010919-023324. Epub 2019 Sep 13. PMID: 31518513; PMCID: PMC7010403. 7. Kelsey Chetnik, Elisa Benedetti, Daniel P Gomari, Annalise Schweickart, Richa Batra, Mustafa Buyukozkan, Zeyu Wang, Matthias Arnold, Jonas Zierer, Karsten Suhre, Jan Krumsiek, maplet: an extensible R toolbox for modular and reproducible metabolomics pipelines, Bioinformatics, Volume 38, Issue 4, February 2022, Pages 1168–1170, https://doi.org/10.1093/bioinformatics/btab741 8. Peng, Roger D. R programming for data science. Victoria, BC, Canada: Leanpub, 2016 9. www.bioconductor.org 10. Kumar, R., Bohra, A., Pandey, A. K., Pandey, M. K., Kumar, A. (2017). Metabolomics for plant improvement: Status and prospects. Frontiers in Plant Science, 8, 1302. https://doi.org/10.3389/fpls.2017.01302