Specmine: an R package for metabolomics and spectral data analysis and mining

Christopher Costa[1], Telma Afonso[1], Sara Cardoso[1], Marcelo Maraschin[2], Miguel Rocha[1]

[1] *CEB - Centre of Biological Engineering, Universidade do Minho, 4710-057 Braga, Portugal*
[2] *Plant Morphogenesis and Biochemistry Laboratory, Federal University of Santa Catarina, 88049-900, Florianopolis, Brazil*
Group: BIOSYSTEMS | Line: Industrial and Food Biotechnology and Bioengineering

In the last years, interest in the field of metabolomics has been growing, materialized by the advances in experimental techniques, growth in available data and novel biological applications. Techniques as Nuclear Magnetic Resonance, Gas or Liquid Chromatography, Mass Spectrometry, and Infrared, UV-visible or Raman spectroscopies have provided extensive datasets that can help in tasks as biological and biomedical discovery, biotechnology and drug development. However, as it happens with other omics data, the analysis of metabolomics datasets provides methodological and computational challenges. Indeed, from the available software tools, none addresses the multiplicity of existing techniques and data analysis tasks.

At the Biosystems group, in collaboration with the UFSC in Brazil, we have developed a novel R package, named *specmine*, which provides a set of methods for metabolomics data analysis, including data loading in different formats, pre-processing, metabolite identification, univariate and multivariate data analysis, machine learning, and feature selection. Importantly, the implemented methods provide adequate support for the analysis of data from diverse experimental techniques, integrating a large set of functions from several R packages in a powerful, yet simple to use environment.

The *specmine* package is available in the CRAN R repository to be installed by any interested user. It has already been used to address data analysis tasks in different scenarios, considering natural products analysis, food and environmental research, tackling challenges as the characterization of bees propolis, cassava post-harvest deterioration or carotenoid contents or the exposure of algae to pollutants as diesel or gasoline.

Currently, we are finalizing the development of *webspecmine*, an online data analysis tool, which aims to provide the features of *specmine* through a web-based interface, making it easier to use by non-informaticians. This site will also prvide tools to enable researchers to deposit their datasets and make them available for the community, boosting data sharing and data analysis reproducibility.

References

[1] Costa, Christopher, Marcelo Maraschin, and Miguel Rocha. "An R package for the integrated analysis of metabolomics and spectral data." Computer methods and programs in biomedicine 129 (2016): 117-124.