MetaR: simple, high-level languages for data analysis with the R ecosystem

Fabien Campagne[1,2,3,*], William ER Digan[1], Manuele Simi[1,2]

[1]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, United States of America; [2]Clinical Translational Science Center, Weill Cornell Medicine, New York, NY, United States of America; [3]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, United States of America. *To whom correspondence should be addressed: fac2003@campagnelab.org

Keywords: Data Analysis, Reproducibility, Docker, Language Workbench Technology

Data analysis tools have become essential to the study of biology. Here, we applied language workbench technology (LWT) to create data analysis languages tailored for biologists with a diverse range of experience: from beginners with no programming experience to expert bioinformaticians and statisticians.

A key novelty of our approach is its ability to blend user interface with scripting in a single platform. This feature helps beginners and experts alike analyze data more productively.

This new approach has several advantages over state of the art approaches currently popular for data analysis: experts can design simplified data analysis languages that require no programming experience, and behave like graphical user interfaces, yet have the advantages of scripting. We report on such a simple language, called MetaR [1], which we have used to teach complete beginners how to call differentially expressed genes and build heatmaps. We found that beginners can complete this task in less than 2 hours with MetaR, when more traditional teaching with R and its packages would require several training sessions (6-24hrs). Furthermore, MetaR seamlessly integrates with docker to enable reproducibility of analyses and simplified R package installations during training sessions.

We used the same approach to develop the first composable R language. A composable language is a language that can be extended with micro-languages. We illustrate this capability with a Biomart micro-language designed to compose with R and help R programmers query Biomart interactively to assemble specific queries to retrieve data, (The same micro-language also composes with MetaR to help beginners query Biomart.)

Our teaching experience suggests that language design with LWT can be a compelling approach for developing intelligent data analysis tools and can accelerate training for common data analysis task. LWT offers an interactive environment with the potential to promote exchanges between beginner and expert data analysts. This talk will provide an introduction to LWT and describe our results with the MetaR platform [1,2].

Software is distributed under the Apache 2.0 license and available at https://github.com/CampagneLaboratory/MetaR and http://metaR.campagnelab.org.

References: [1] Fabien Campagne, William ER Digan, Manuele Simi. MetaR: simple, high-level languages for data analysis with the R ecosystem. BioRxiv doi: http://dx.doi.org/10.1101/030254 [2] Documentation http://tinyurl.com/zx9wvjw