Automating biostatistics workflows for bench scientists using R-based web-tools

Jeff Skinner, Vivek Gopalan, Jason Barnett, Yentram Huyen[*]

Bioinformatics and Computational Biosciences Branch (BCBB)
Office of Cyber Infrastructure and Computational Biology (OCICB)
National Institute of Allergy and Infectious Diseases (NIAID)
National Institutes of Health (NIH)
[*]Contact author: huyeny@niaid.nih.gov

Keywords: Web-tools, Computational Biology, Curve-fitting, Structural Biology, NIH

Biological data can be complex, so bench scientists often develop complicated workflows to process, analyze and present their data. Often biological data will be output from a laboratory instrument as a text data file, then a researcher will spend hours processing the data in MS Excel®, performing analyses in a commercial statistics software package or a custom built script written in FORTRAN or perl, before processing the results further back in Excel and finally producing a report in MS Word® or PowerPoint®. These complicated workflows are tedious and time consuming; they introduce multiple opportunities for error and they can be difficult for future researchers to reproduce. It may be easy to replicate these workflows in R, but its steep learning curve prevents many bench scientists from using R scripts that might simplify their analyses. Statisticians and R programmers need to provide more intuitive user interfaces before their R scripts can be widely adopted by biologists. We present results from two web-based tools, which use R to reproduce critical analysis workflows while providing bench scientists with a simple point-and-click GUI front-end. The Dose-Response Analysis Pipeline (DRAP) allows immunologists to apply curve-fitting analyses to multiple dose-response experiments conducted on one or more 96-well plates. Logistic curve-fit results can be compared among several groups or factors distributed within or among the 96-well plates. Final results are presented in an interactive PDF report with high-resolution images. The DRAP workflow was able to analyze approximately 2000 plates in 30 minutes, which had taken more than 200 hours to analyze by hand (`http://exon.niaid.nih.gov/DRAP`). The Hydrogen Exchange with Normalized Assessment of Maximum Entropy (HDXNAME) workflow allows structural biologists to compute protein flexibility estimates called protection factors from hydrogen exchange experiment data using Maximum Entropy Methods (MEM). This single workflow replaces several cut-and-paste procedures between hard-to-find Excel templates and a custom software application written in MS BASIC. The final result includes a statistical summary comparing two conformational states of the protein and an image of the protein with protection factors mapped on the protein surface (`http://exon.niaid.nih.gov/HDX_NAME`). These two examples show how R programming and web-site design can be used to create meaningful custom applications that will be widely used and shared among biology researchers.

References

JM Sa, O Twu, K Hayton, S Reyes, MP Fay, P Ringwald and TE Wellems (2009). Geographical patterns of *Plasmodium falciparum* drug resistance distinguished by differential responses to amodiaquine and chloroquine. *PNAS*, 106(45): 18883–18889.

L Kong, C Huang, SJ Coales, KS Molnar, J Skinner, Y Hamuro and PD Kwong (2010). Local conformational stability of HIV-1 gp120 in unliganded and CD4-bound states as defined by amide hydrogen/deuterium exchange. *in preparation*