biotmle: Targeted Learning for Biomarker Discovery

Nima S. Hejazi¹, Weixin Cai¹, and Alan E. Hubbard¹

¹ Division of Biostatistics, University of California, Berkeley

Summary

The biotmle package provides an implementation of a biomarker discovery methodology based on targeted minimum loss-Based estimation (TMLE) (van der Laan and Rose 2011) and a generalization of the moderated t-statistic of (Smyth 2004), designed for use with biological sequencing data (e.g., microarrays, RNA-seq). The statistical approach made available in this package relies on the use of TMLE to rigorously evaluate the association between a set of potential biomarkers and another variable of interest while adjusting for potential confounding from another set of user-specified covariates. The implementation is in the form of a package for the R language for statistical computing (R Core Team 2017).

There are two principal ways in which the biomarker discovery techniques in the biotmle R package can be used: to evaluate the association between (1) a phenotypic measure (say, environmental exposure) and a biomarker of interest, and (2) an outcome of interest (e.g., survival status at a given time) and a biomarker measurement, both while controlling for background covariates (e.g., BMI, age). By using an estimation procedure based on TMLE, the package produces results based on the Average Treatment Effect (ATE), a statistical parameter with a well-studied causal interpretation (see van der Laan and Rose (2011) for extended discussions), making the biotmle R package well-suited for applications in bioinformatics, epidemiology, and genomics.

After adjusting our data set to be consistent with the expect input format – please consult the vignette accompanying the R package for details – we would call the principal function of this R package: biomarkertmle.

We would perform a moderated test on the output of the biomarkertmle function using the function modtest_ic.

While the principal table of results produced by this R package matches those produced by the well-known limma R package (Smyth 2005), there are also several plot methods made available for the bioTMLE S4 class – subclassed from the popular SummarizedExperiment class – introduced by this package (Huber et al. 2015). For illustrative purposes, we demonstrate the output of two such functions on anonymized experimental data below:
Figure 1: Heatmap visualizing the Average Treatment Effect contribution of a change in exposure to each biomarker of interest
Figure 2: Volcano plot visualizing the log fold change in the Average Treatment Effect against the raw p-value from the moderated t-test performed on each biomarker
References


