TARGETED MACHINE LEARNING FOR UNDERSTANDING HIV RESISTANCE TO NEUTRALIZING ANTIBODIES

Yutong Jin[1], David Benkeser[1]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA

EMORY ROLLINS SCHOOL OF PUBLIC HEALTH | Department of Biostatistics and Bioinformatics

Abstract

Motivation: Human immunodeficiency virus (HIV) infection, caused by the virus that attacks and destroys human immune system, can gradually advance to Acquired Immune Deficiency Syndrome (AIDS) if not treated. Protective antibodies have been successfully isolated from infected individuals to target HIV-1 envelope spike, and prevention trials involving passive infusion with monoclonal antibodies are currently being conducted. A potential challenge for monoclonal antibody regimens is the large genetic diversity of the HIV envelope, which could allow some viruses to escape neutralizing actions of antibodies. It is therefore of interest to develop methodology to identify resistant envelope sequences. This information could help guide the selection of antibodies to be included in monoclonal antibody regimens that include multiple antibodies.

Method: e proposed a machine learning-based method for identifying amino acid (AA) residues along the envelope protein that are causally predictive of resistance to antibody neutralization using in vitro neutralization assays.

Results: We demonstrate via simulation that the approach enjoys important statistical benefits over existing approaches. We apply the approach to the Compile, Analyze and Tally Nab Panels (CATNAP) database to identify 24 AA positions that are potentially causally related to the resistance to neutralization by the VRC01 antibody.

Introduction

Background:

To reduce incidence further, there is great demand for the development of a preventive HIV vaccine. An effective HIV-1 vaccine will likely require an antigen that induces production of antibodies capable of neutralizing viral strains with genetically diverse envelope proteins (Gaschen *et al.*, 2002). Broadly neutralizing antibodies (bNAbs) are considered a highly promising means of HIV prevention, but, as with vaccines, the key obstacle is selection of antibodies that are capable of neutralizing genetically diverse HIV strains. This challenge has led to the idea that it may be necessary to use a cocktail of bNAbs to increase breadth of coverage.

Challenges:

- How to identify potentially influential sites of a particular bNAb to reduce the impact of multiple testing adjustments.
- How to select a combination of bNAbs that are likely to provide sufficient neutralization breadth
- Positivity assumption violation:
  - structural constraints in the Envelope protein
  - high-dimensional nature

} Inflated type I errors

We propose a procedure that seeks to overcome these difficulties using a random forest-based, outcome-adpative, collaborative targeted minimum loss-based estimation (CTMLE) approach.

Material and Methods

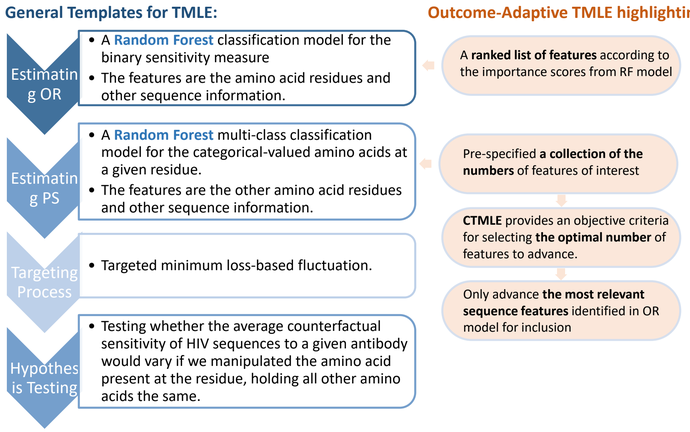

Figure 1: Flow chart for the basic and outcome-adaptive TMLE approaches

Simulation Study

We conducted different Monte-Carlo simulation studies to compare the performance of proposed TMLE and CTMLE estimators with reduced feature dimension in PS model to the TMLE estimator with all features included. Monte-Carlo data sets were generated as follow:

Sample size: **500 / 1000**

The number of features (Residues): **200**

The number of levels for each features: **4**

Feature structure: **AR-1 correlation** with $\boldsymbol{\rho = 0.75}$ (moderate correlated)

True signals: $\boldsymbol{AA_{37}, AA_{87}, AA_{94}, AA_{135}, AA_{151}}$

A collection of numbers of selected features in Propensity Score model: **5, 10, 50, 100, 200**

Three covariates were selected for demonstration:

(1) $\boldsymbol{AA_{10}}$, ***non signal uncorrelated*** with any true signals;

(2) $\boldsymbol{AA_{85}}$, ***non signal highly-correlated*** with certain true signals;

(3) $\boldsymbol{AA_{87}}$, ***true signal***.

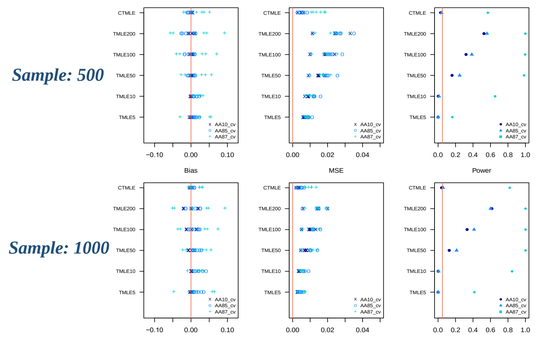

Figure 2: Bias, mean-squared error (MSE), type I error rate for non-signals and power for true signals

Manually setting of the cutoff value was unreliable and a method of optimal number selection was of great demand. As expected, CTMLE offered some benefits in terms of appropriately controlling type I error; it achieved as well as the best model among all simulated scenarios.

Data analyses: CATNAP data VRC01

VRC01 is a human bNAbs targeting at CD4-binding site of HIV envelope glycoprotein, which was isolated from human in 2010 and has been assessed for HIV-1 prevention potency (Gilbert *et al.*, 2017). Compile, Analyze and Tally NAb Panels (**CATNAP**) database is a new comprehensive platform integrating the information of neutralization antibodies and envelope residues with **828** observations (Yoon *et al.*, 2015). This database consists of a binary result of VRC01 sensitivity, defined as a inhibitory concentration with 50% reduction, **784** site-specific envelope AA sequences and several other demographic records of virus, including the geographic origin, subtype, and size. After applying the screening steps, there are **328** AA residues remaining for further investigation. The results were summarized in **Figure 3**.

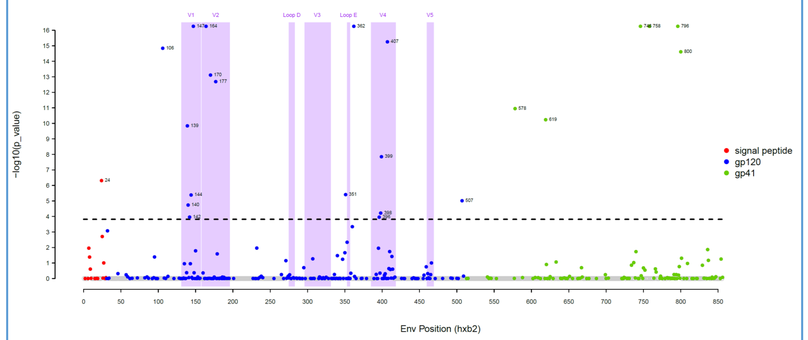

Figure 3: Amino acids designations with annotations of gp120 functional regions

It demonstrated that, AA mutations at each site were differently associated with neutralization resistance among 23 (7%) predictive residues.

Discussion

Strength:

- Excluding irrelevant features for PS model could efficiently reduce type I error for non-signals no matter barely or highly related to the true significant AA residues.
- When sample size is large enough, CTMLE approach could achieve a high power with a better controlled type I error.

Generalization:

- This method can be extended from VRC01 to other bNAbs
- The Feature importance can be obtained from other algorithms
- The current pairwise comparison can be extended to include some family-wise error rate control methods

Reference

Gaschen, B. *et al.* (2002). Diversity considerations in hiv-1 vaccine selection. *Science*, **296**(5577), 2354–2360.

Gilbert, P. B. *et al.* (2017). Basis and statistical design of the passive hiv-1 antibody mediated prevention (amp) test-of-concept efficacy trials. *Statistical communications in infectious diseases*, **9**(1).

Yoon, H. *et al.* (2015). Catnap: a tool to compile, analyze and tally neutralizing antibody panels. *Nucleic acids research*, **43**(W1), W213–W219.

Contact

Yutong Jin email: yutong.jin@emory.edu