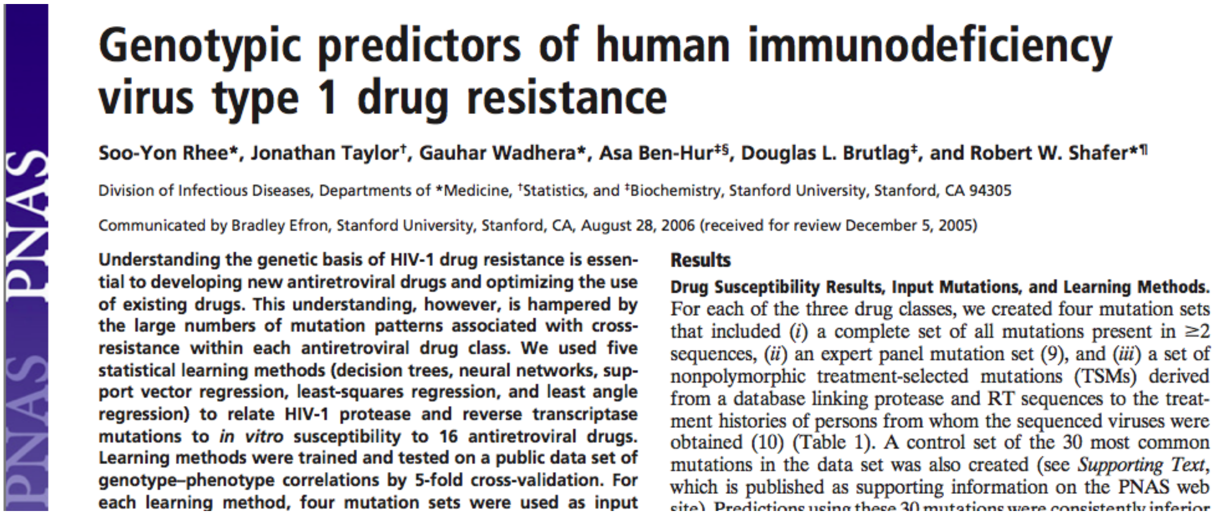
STAT460 – Homework 4
Due: Feb. 18 at the start of class.

A major issue with antiretroviral drugs is the mutation of the virus' genes. Because of its high rate of replication (10^9 to 10^{10} virus per person per day) and error-prone polymerase[1], HIV can easily develop mutations that alter susceptibility to antiretroviral drugs. The emergence of resistance to one or more antiretroviral drugs is one of the more common reasons for therapeutic failure in the treatment of HIV.

In the following paper[2], a sample of *in vitro*[3] HIV viruses were grown and exposed to a particular antiretroviral therapy. The susceptibility of the virus to treatment and the number of genetic mutations of each virus were recorded.



Genotypic predictors of human immunodeficiency virus type 1 drug resistance

Soo-Yon Rhee[*], Jonathan Taylor[†], Gauhar Wadhera[*], Asa Ben-Hur[‡§], Douglas L. Brutlag[‡], and Robert W. Shafer[*¶]

Division of Infectious Diseases, Departments of [*]Medicine, [†]Statistics, and [‡]Biochemistry, Stanford University, Stanford, CA 94305

Communicated by Bradley Efron, Stanford University, Stanford, CA, August 28, 2006 (received for review December 5, 2005)

Understanding the genetic basis of HIV-1 drug resistance is essential to developing new antiretroviral drugs and optimizing the use of existing drugs. This understanding, however, is hampered by the large numbers of mutation patterns associated with cross-resistance within each antiretroviral drug class. We used five statistical learning methods (decision trees, neural networks, support vector regression, least-squares regression, and least angle regression) to relate HIV-1 protease and reverse transcriptase mutations to *in vitro* susceptibility to 16 antiretroviral drugs. Learning methods were trained and tested on a public data set of genotype–phenotype correlations by 5-fold cross-validation. For each learning method, four mutation sets were used as input

Results

Drug Susceptibility Results, Input Mutations, and Learning Methods. For each of the three drug classes, we created four mutation sets that included (*i*) a complete set of all mutations present in ≥2 sequences, (*ii*) an expert panel mutation set (9), and (*iii*) a set of nonpolymorphic treatment-selected mutations (TSMs) derived from a database linking protease and RT sequences to the treatment histories of persons from whom the sequenced viruses were obtained (10) (Table 1). A control set of the 30 most common mutations in the data set was also created (see *Supporting Text*, which is published as supporting information on the PNAS web site). Predictions using these 30 mutations were consistently inferior

1. Load the data set `hiv.rda` and create

   ```
   X = hiv.train$x
   Y = hiv.train$y
   ```

   What would be n and p in this problem? What are the covariates in this problem? What are the observations? What is the response? **Note:** Attempt to answer this question before moving on to the rest of the questions.

2. Consider the design matrix $\mathbb{X}$. It is composed of 0's and 1's, with a 1 indicating a mutation in a particular gene. Run

   ```
   table(X)
   ```

   What results do you get? What does this indicate?

3. The response the log transformed susceptibility of a virus to the considered treatment, with large values indicating the virus is resistant (that is, not susceptible). Run

[1] An enzyme that 'stitches' back together DNA or RNA after replication

[2] The entire paper is on the website. Try to see what you can get out of it.

[3] Latin for *in glass*, sometimes known colloquially as a test tube

```
hist(Y)
```

What plot did you just create? What does this indicate?

4. We may have (at least) two goals with a data set such as this: inference or prediction. An inferential question would be: can we find some genes whose mutation seems to be most related to viral susceptibility? A prediction question would be: can we make a model that would predict whether this therapy would be efficacious, given a virus with a set of genetic mutations.

   (a) Let's do model selection, which can address either of these goals.

      i. Try to find either the best subset solution for this problem. Discuss any problems or findings you discover. In particular, how many possible models are there?

      ii. Now do forward selection with `regsubsets`. Report the selected covariates using `bic` as the criterion.

   (b) Now, let's do ridge regression, which only addresses prediction. Using the package `glmnet`, find the minimum CV ridge solution and report its CV estimate of the prediction risk.

   Note: There is no need to report the p coefficient estimates from the ridge solution.