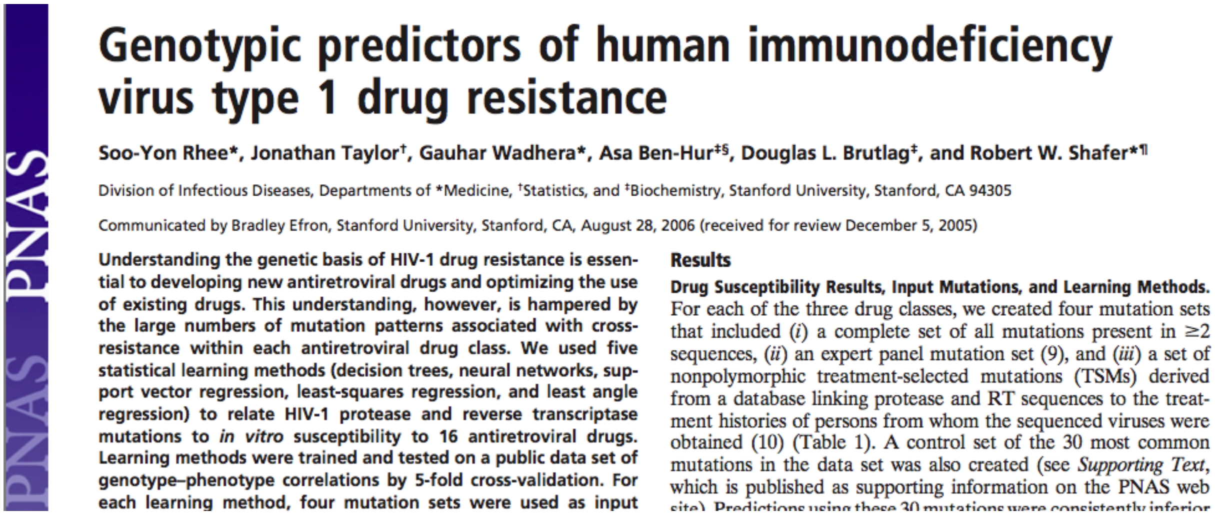
STAT460 – Homework 5
Due: Feb. 25 at the start of class.

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.



PNAS PNAS

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

Anytime I ask for 'test set prediction error' for a method, use the following

```
X_0 = hiv.test$x
Y_0 = hiv.test$y

Y.hat = ###### prediction of method on X_0 #######

print(mean((Yhat - Y_0)**2))
```

Let's look at comparing the lasso with CV, elastic net, refitted lasso, and scaled sparse regression.

1. Find the test set prediction error for lasso with CV. Compare the CV estimate of the risk to the test set prediction.

2. Find the test set prediction error for elastic net with

[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

   (a) $\alpha = .5$
   (b) $\alpha = .9$

3. Find the test set prediction error for refitted lasso
4. Find the test set prediction error for SSR